Deep Supervision of the Vocal Tract Shape for Articulatory Synthesis of Speech

Vinicius Ribeiro

Université de Lorraine

Ecole Doctorale IAEM

Thèse de Doctorat

Specialité **Informatique**

présentée et soutenue publiquement le 5 décembre 2023

par

Vinícius de Paulo Souza Ribeiro

Deep Supervision of the Vocal Tract Shape for Articulatory Synthesis of Speech

sous la direction de **Yves Laprie**

Composition du jury

Président du Jury:	Anne Boyer	Université de Lorraine
Rapporteurs:	Damien Lolive	ENSSAT Lannion, Université de Rennes
	Antoine Serrurier	RWTH University Aachen
Examinateurs:	Eduardo Valle	Universidade Estadual de Campinas
	Alice Turk	University of Edinburgh
	Pierre-André Vuissoz	Université de Lorraine
Directeur de Thèse:	Yves Laprie	Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA),

UMR 7503 Équipe Multispeech, 54506, Vandœuvre-lès-Nancy, France

To my father, mother, and sister.

If I have seen further, it is by standing upon the shoulders of giants.

Isaac Newton

Acknowledgements

The epigraph of this thesis is a famous quote by Isaac Newton. In Newton's vision, the giants, such as Galileo and Copernicus, were the scientists that preceded him. Of course, the scientists that preceded me enormously impacted my work. However, I'll defer their contributions to other parts of this text. Here I refer to giants outside the history books without published papers or Nobel prizes (yet). I refer to three precisely: my father, Wanius Ribeiro; my mother, Raquel de Souza Ribeiro; and my sister, Nicole de Souza Ribeiro. If I have seen further, succeeded in my experiments, and finished writing my thesis, it was by standing on their shoulders. I could not be more proud and thankful for the people that raised me. If only I could be a tiny fraction of who they are, I could consider myself successful. A thousand diplomas would not fulfill me more than that. I want to thank my parents and sister for being with me through the most challenging times. I will never walk alone as long as I carry them in my heart.

I want to thank my thesis supervisor, Yves Laprie, whose guidance and support were paramount to my work. Yves is one of the kindest people I have met and the most supportive supervisor I could ever wish for. I moved to France during one of the roughest times of the COVID-19 pandemic. Yves was always concerned about my well-being and personally involved in reducing the friction in my adaptation. Yves is an insightful, ethical, and responsible scientist who allowed me to pursue my research goals and proved that extracting the best from people without losing kindness is possible. This thesis would never be possible without him, and I will always be thankful to him.

The Centre Hospitalier Régional Universitaire (CHRU) de Nancy was an essential partner in this research. I want to thank Pierre-André Vuissoz, Karyna Isaieva, and Justine Leclere, who worked with me during the entire period of my doctorate and were the co-authors of many of my publications during these years. The partnership with CHRU allowed us to collect high-quality data of utmost importance in my research. Pierre-André, Karyna, and Justine provided a completely different point of view from my background. Also, they provided me with valuable research experience, advice, and knowledge in medical imaging and anatomy.

The Multispeech team welcomed me right after I arrived in France. Together we developed high-quality research, shared interesting ideas, grew as scientists, and shared terrific relaxation moments. I want to thank all the people who were part of this team during my stay. In particular, I want to thank Ajinkya Kulkarni, Sandipana Dowerah, Tulika Bose, and Marina Kréme for their immense friendship and the great times we spent together. Ajinkya and Sandipana were the first friends I made in the laboratory. Tulika and Marina were always supportive and great teammates. I hope I can carry these relationships for the rest of my life. Moreover, I

want to thank Denis Jouvet and Slim Ouni for leading our team during my stay. Their help and leadership were extraordinary in achieving relevant results. Likewise, I want to thank all the Loria's staff, including but not limited to the administrative team, human resources, restaurant personnel and receptionists, for their hard work. A substantial amount of the scientific contributions made in the lab is thanks to the people that work on the back office creating this amazing work environment.

During my life, I made many friends that I will always carry with me. I want to thank all the friends and family who impacted my life. Moreover, I want to thank a few people especially. Omar Darwiche Domingues was one of the most important people in my Ph.D. Omar is a brilliant scientist and one of my best friends. He was always supportive during these years, helping me to move through difficult moments, giving his opinion about my papers, and advising me to grow as a better researcher.

Distance usually tears people apart. It is the painful truth. With my friends, however, it was the opposite. Daniel Melges, Paulo Azevedo, Edgar Berg, Carolina Abrahão, and Jessica Castro were always there for me. When the distance was enormous, they managed to be close and took the time to call or send a message to check on me. Their presence was paramount in helping me overcome the homesickness. I feel so fortunate and thankful for calling them all best friends.

Nancy was my home for the last three years. The city brought me the most valuable thing I could expect. Christian Pereira, Barbara Moissa, Guilherme Alves, and Jessica Aragão are a small piece of Brazil in Europe for me. We shared unforgettable moments playing board games, cooking barbecue, or just talking, laughing, and enjoying each other's presence. I did not expect to connect with anyone so fast when I left my country. I could not have been more wrong.

During my Ph.D., I had the fantastic opportunity to join Meta AI, in New York City, USA, for a sixteen-week internship, where I participated in a challenging problem in a new domain, learning from some of the best researchers and engineers in the world. I want to thank all my colleagues on the AI Speech Team at Meta. More specifically, I want to thank Yiteng Huang, who mentored and managed me, providing valuable and actionable feedback and advice I will carry throughout my career. Likewise, I want to thank my colleagues Yuan Shangguan, Zhaojun Yang, Li Wan, and Ming Sun, who provided me with technical support, research insights, and guidance during my work. Also, I would like to thank Philipp Klumpp, who joined the internship program with me and became a close friend and an outstanding chess tutor.

The Ph.D. journey ends in the defense, and a new one starts right after. Therefore, last, but not least, I want to thank the members of the jury for participating in this important moment,

providing great comments, advice and proportioning a fantastic scientific debate. I am looking forward to live the opportunities and face the challenges the future has to offer.

Abstract

Speech is a dynamic and non-stationary process that requires the interaction of several vocal tract articulators. The context in which a phoneme is articulated strongly influences its production, a phenomenon known as coarticulation. Articulatory speech synthesis and its counterpart, acoustic-to-articulatory inversion, hold many potential applications, such as L2 learning and speech therapy design. Moreover, these models are helpful for speech synthesis and automatic speech recognition since they create a link to the speech production process.

Modeling speech articulations presents challenges such as coarticulation, non-uniqueness, and speaker normalization. Historically, the research focused on geometrical, mathematical, and statistical models to describe speech dynamics. Nevertheless, developing such models faces the difficulty of obtaining relevant articulatory data from actual speakers. Since the vocal tract is not observable from the outside, various invasive and non-invasive methods have been used to collect these data, including flesh point tracking and medical imaging. The first attempts to extract articulatory data used X-rays, but it was abandoned due to exposure to ionizing radiation. Then, electromagnetic articulography rapidly grew in popularity due to the high sampling rate and the low cost compared to the alternatives. More recently, real-time magnetic resonance imaging (RT-MRI) has been the preferred acquisition method due to the visibility of the vocal tract from the glottis to the lips.

This thesis explores the synthesis of speech articulation movements corresponding to a sequence of phonemes. The primary objective is to design a model that predicts the temporal evolution of the vocal tract shape for each phoneme in the input sequence. Nevertheless, developing a realistic temporal model of the vocal tract is challenging. We split the problem into three contributions.

The first is obtaining the vocal tract profile from the RT-MRI films by developing a robust method for segmenting vocal tract articulations. The second contribution is to build an articulatory model that predicts the vocal tract shape for any phonetic input in French. The challenges are learning coarticulation and enforcing the places of articulation and articulatory movements that lead to the expected acoustics. The third contribution is the evaluation of the predicted shapes. We propose to quantify phonetic information with the aid of phoneme recognition. We measure the phonetic information retained by the mid-sagittal contours and that reproduced by the vocal tract shape synthesizer using the phoneme error rate and the recognizer's internal representations.

This thesis points to significant directions in speech articulation synthesis. We observe that model-free synthesis, without an articulatory model, leads to the best and most natural results.

Nevertheless, using an intermediate articulatory model permits the introduction of relevant phonetic knowledge into the model. Finally, we open a new direction to evaluate articulatory models through their phonetic representations.

Résumé

La parole est un processus dynamique et non stationnaire qui nécessite l'interaction de plusieurs articulateurs du conduit vocal. Le contexte dans lequel est articulé un phonème influence très fortement sa production, ce phénomène est connu sous le nom de coarticulation. La synthèse articulatoire de la parole et son homologue, l'inversion acoustique-articulatoire, ont de nombreuses applications potentielles, telles que l'apprentissage des langues étrangères et la conception d'approches de remédiation de la production de la parole. De plus, ces modèles sont utiles pour la recherche en synthèse et en reconnaissance automatique de la parole parce qu'ils font le lien avec le processus de production de la parole.

La modélisation des articulations de la parole présente des défis tels que la coarticulation, la non-unicité, et la normalisation du locuteur. Historiquement la recherche s'est concentrée sur les modèles géométriques, mathématiques et statistiques pour décrire la dynamique de la parole. Néanmoins, le développement de tels modèles est confronté à la difficulté d'obtenir des données articulatoires pertinentes auprès de locuteurs réels. Le conduit vocal n'étant pas observable de l'extérieur, diverses méthodes invasives et non invasives ont été utilisées pour collecter ces données, notamment le suivi de capteurs collés sur les articulateurs et l'imagerie médicale. Les premières techniques d'extraction de données articulatoires ont utilisé des rayons X, mais cette technique a été abandonnée en raison de l'exposition aux rayonnements ionisants. Ensuite, l'articulographie électromagnétique a rapidement gagné en popularité en raison de sa fréquence d'échantillonnage élevée et de son faible coût par rapport aux autres techniques. Plus récemment, l'imagerie par résonance magnétique en temps réel (RT-MRI) est devenue la méthode d'acquisition privilégiée en raison de la visibilité de tout le conduit vocal depuis la glotte jusqu'aux lèvres.

Cette thèse explore la synthèse des mouvements articulatoires de la parole correspondant à une séquence de phonèmes. L'objectif principal est de concevoir un modèle qui prédit l'évolution temporelle de la forme du conduit vocal pour chaque phonème de la séquence d'entrée. Néanmoins, le développement d'un modèle temporel réaliste du conduit vocal est un défi. Nous avons décomposé le problème en trois contributions.

La première consiste à obtenir le profil du conduit vocal à partir des films d'IRM temps réel en développant une méthode robuste de segmentation des articulations du conduit. La deuxième contribution consiste à construire un modèle articulatoire qui prédit la forme du conduit vocal pour toute entrée phonétique en français. Les défis sont d'apprendre la coarticulation et d'imposer les lieux d'articulation et les mouvements articulatoires qui conduisent à l'acoustique attendue. La troisième contribution est l'évaluation des formes prédites par le modèle. Nous

proposons de quantifier l'information phonétique à l'aide de la reconnaissance automatique de phonèmes. Nous mesurons l'information phonétique capturée par les contours médiosagittaux et celle reproduite par le synthétiseur de la forme du conduit vocal en utilisant le taux d'erreur phonétique et les représentations internes du reconnaisseur.

Cette thèse ouvre des pistes importantes pour la synthèse articulatoire de la parole. Nous avons observé que la synthèse model-free, c'est-à-dire sans modèle articulatoire, conduit aux meilleurs résultats et aux plus naturels. Néanmoins, l'utilisation d'un modèle articulatoire intermédiaire permet d'introduire des connaissances phonétiques pertinentes dans le modèle. Enfin, ce travail ouvre une nouvelle piste de recherche pour évaluer les modèles articulatoires à travers leur représentation phonétique.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

> Words are, in my not so humble opinion, our most inexhaustible source of magic, capable of both influencing injury, and remediating it.
>
> *Albus Dumbledore*
> *J. K. Rowling*

1 Overview

The ability to express meaningful sounds and modulate our voices is essential for human communication. The development of an unprecedented form of communication was fundamental in the evolution of the *homo sapiens* and its success in conquering its environments against other human and non-human species. Yuval Noah Harari, in his best-selling book "*Sapiens: A Brief History of the Humankind*" [1], explains how the Cognitive Revolution brought to humans the ability to cooperate in large groups through a complex language. The many attributes of the new language gave the *homo sapiens* a large competitive advantage. It enabled detailed explanations of complex events that occurred in distant times, a necessary ability for survival, and permitted humans to negotiate with its counterparts, essential in politics and general business. But most importantly, Harari argues that the main contribution of the Cognitive Revolution was the ability to think and discuss about the unreal and to create the myths that bounds today's society. All of these factors were only possible with the development of language, but most specifically, spoken language.

The contributions of speech in the evolution of the *homo sapiens* are paramount. Language skills are mostly learned rather than innate. From infancy, newborns rapidly discover how to use their premature voice to express their needs. Babies cry when they are hungry or in

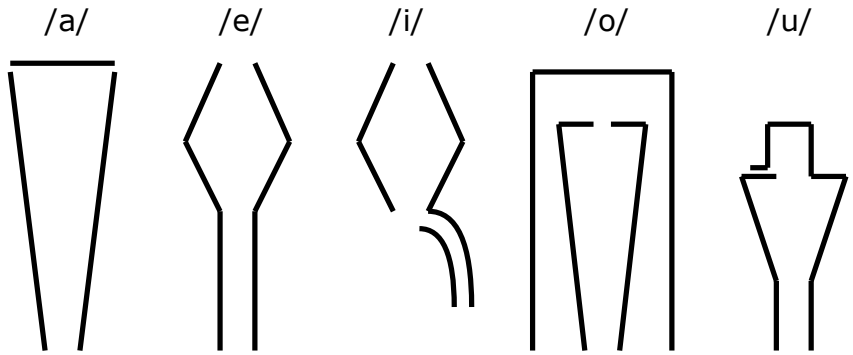


Figure 1.1: Kratzenstein tubes for the five vowels.

pain. With age, humans obtain a greater control of their vocal tract apparatus, being more capable to convert thoughts into spoken language. Karthikeyan et al. [2] quantified the impact of articulatory fluency on socialization, suggesting that men that articulate phonemes more distinctly are seen as more attractive by women for long-term relationships and are perceived by other men as having higher prestige. It is not a surprise then that the dynamics of speech production have long fascinated scientists.

One of the earliest pioneers in this field was Christian Gottlieb Kratzenstein. In 1779, Kratzenstein made the first attempts at what we now call speech synthesis. At that time, the physical mechanisms of sound waves propagation were well understood, but the details of how speech arises from the air flow in the vocal folds and vocal tract were still a mystery. Leaonard Euler suggested that a musical instrument capable of reproducing the five vowels /a, e, i, o, u/ would be possible, and the Academia of Saint Petersburg offered a prize for the first person to achieve this.

Kratzenstein, who was already investigating the topic, won the prize with his design of a set of organs that correspond to the vocal tract resonating cavities for the five vowels (Figure 1.1). When excited by a free reed, these organs could reproduce their respective sounds. His work, titled *"Tentamen resolvendi problema"*, was published in 1781 [3]. Unfortunately, the original tubes were damaged and later lost, but there have been attempts to reconstruct them from historical notes [4].

At the same time, Wolfgang von Kempelen achieved great fame with one of his inventions. Kempelen's speaking machine was a complex device that consisted of a number of interconnected tubes, chambers, and valves. When air is blown into the machine, it would cause propagation of sound waves inside the tubes. The shapes of the tubes and chambers could be adjusted to different vowels and consonants.

Kempelen's machine was a sensation when it was first unveiled in 1791 and detailed in his book *"Mechanismus der menschlichen Sprache"* (The Mechanisms of Human Speech). It was

(a) Original Kempelen's speaking machine.

(b) Replica constructed by the Deutsches Museum.

Figure 1.2: Kempelen's speaking machine. Images reproduced from Google Arts & Culture [5].

exhibited all over Europe, and people were amazed by its ability to produce human speech. However, the machine was difficult to operate, and it could only produce a limited range of sounds. Despite its drawbacks, Kempelen's machine was a significant advance in the development of speech synthesis technology and it inspired other inventors to develop their own machines. Figure 1.2 shows the original design proposed by Von Kempelen and a replica constructed by the Deutsches Museum.

Kratzenstein's and Kempelen's works were a significant advance in the understanding of speech production. Their research continues to be of interest to scientists and engineers today, and it provides valuable insights into the mechanisms of human communication. Professor Kratzenstein and Von Kempelen would be amazed by the advances in speech synthesis technology since their time. Their early attempts at speech synthesis were crude by today's standards, but they laid the foundation for the development of more sophisticated methods.

The first computational attempts to synthesize human speech were based on the mechanics of wave propagation in the vocal tract. This is because speech is produced by the air flow from the lungs that excites the vocal folds when they are sufficiently adducted and with the appropriate tension, originating the sound's fundamental frequency (`f0`). Filtering from the vocal tract cavities creates the formant frequencies. By understanding how these sound waves are produced, it is possible to recreate them artificially.

By the late 1990s to the early 2000s, speech synthesis was still a very rudimentary field. On the one hand, the most accepted approach in industry was concatenative speech synthesis due to the simplicity and reasonable quality [6]. On the other hand, the scientific community was more engaged into articulatory speech synthesis due to the flexibility and the possibility to understand in deep the mechanisms of speech production [7].

With time, the literature moved towards direct approaches to speech synthesis, which would

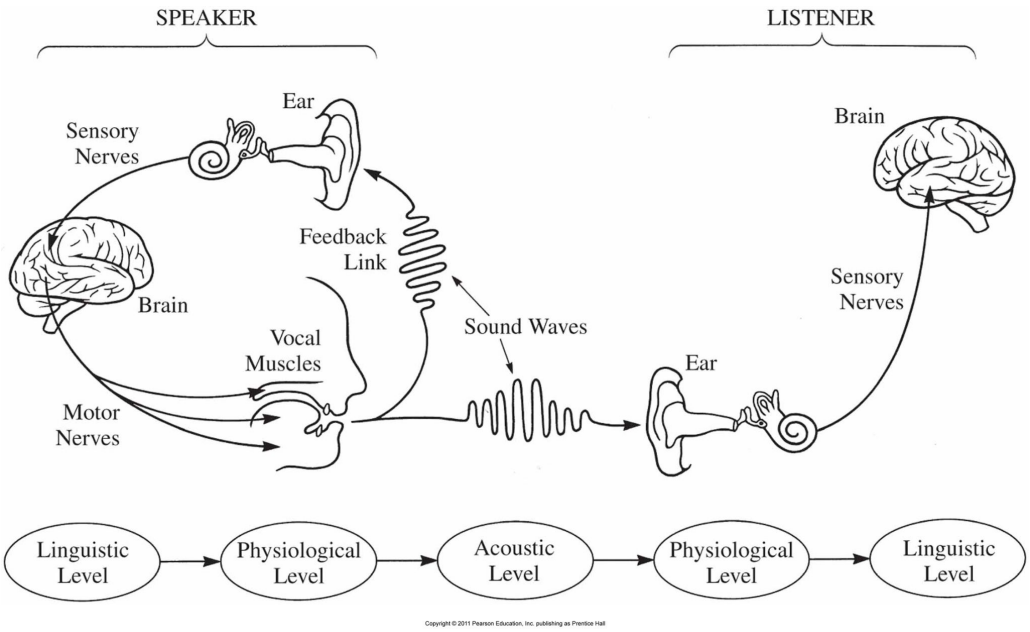


Figure 1.3: The Speech Chain. Originally from Denes and Pinson [10]. Reproduced from Narayanan [11].

not require expensive modeling to generate speech samples. In the first decade of 2000s, Hidden Markov Models [8] started raising attention by providing a statistical view of the speech production process. However, the second decade of 2000s was marked by an enormous growth of deep learning after a long time of disinterest by the scientific community, a period known as the AI winter. Deep neural networks would later dominate the complex data processing research. After that, most of the traditional methods to signal processing started to drop in popularity, and the speech synthesis research concentrated in the use of deep neural networks, which nowadays produce the most realistic speech samples in the literature [9].

Still, articulatory speech synthesis holds its space in research as it has many applications in speech production. In fact, for many voice applications, the physical dynamics of the sound wave propagation in the human vocal tract is irrelevant. However, for understanding how these processes happen from the theoretical point of view, speech therapy research, and language education, these ideas are fundamental. The questions related to these matters are among the main objectives of this thesis.

2 Motivation

Speech communication is usually taken for granted in our daily lives. When speaking, we are unaware of its complexity as it happens automatically. The time difference between thinking and

speaking is a fraction of a second. However, the entire speech chain is elaborated (Figure 1.3). Speaking requires orchestrating different mental, physiological, physical, and social processes. The information is encoded at multiple levels as it progresses from neural-cognitive information in the speaker's brain to sound waves and then back to electrical signals in the listener's brain. Socio-behavioral aspects, such as tone, intent, emotions, demographic traits, and social state, also profoundly influence how others perceive the messages.

More than exploring speech acoustics, multimodal methods permit understanding the structure and function of the vocal tract instrument. Articulatory data enables the comprehension of how the vocal tract transforms the signal encoded by our brain into sounds propagating through the air. In this regard, acquiring and processing articulatory data becomes a central point in speech research. Several techniques permit observing speech production, such as ultrasound for profiling the tongue, electromagnetic articulography for flesh points tracking, electropalatography for tongue-palatal interaction, and X-ray and RT-MRI for complete vocal tract observation.

Articulatory data are essential for detailing the vocal tract morphology and studying and modeling speech communication and their applications extend to many domains, such as healthcare. Hagedorn et al. [12] characterized vocal tract articulations in apraxic speech. Apraxia affects the appropriate selection and temporal coordination of vocal tract gestures; however, these actions may not impact auditory perception. Then, using articulatory features obtained with RT-MRI facilitates the diagnosis. Moreover, Hagedorn et al. [13, 14] studied vocal tract shaping and compensatory strategies in glossectomy patients – when the tongue is partially removed due to cancer. The follow-up research includes using articulatory data to enhance and accelerate therapy.

Furthermore, articulatory speech research extensively impacts conventional speech research (based only on the acoustic signal). Li et al. [15] improved speaker verification systems by including articulatory information (from articulatory inversion) to the acoustic features. Srinivasan et al. [16] explores the robustness of automatic speech recognition systems for neutral and whispered speech and shows that articulatory information is helpful in both scenarios.

These works show how various areas benefit from the characterization and synthesis of vocal tract shape during speech production, and advances to articulatory models have the potential to benefit society in multiple fields.

3 Thesis Objectives

Given the historical footprints and the motivations for our research, this thesis aims to explore and develop deep learning models for articulatory synthesis of speech. More specifically, our

contributions target three main topics: articulatory data processing, articulatory modeling, and model evaluation.

Articulatory Data Processing

Articulatory data acquisition and processing are essential to machine learning. Our primary data source is real-time magnetic resonance imaging (RT-MRI), which is challenging since the images alone are insufficient for speech articulation synthesis. Our first challenge was processing these images to obtain the vocal tract articulators' contours for the entire acquisition. The literature provides multiple approaches to the problem; however, we identified prohibitive gaps for advancing our work such as the difficulty in accessing research data, lack of a gold standard for data annotation, and unavailability of source code. Thus, our first research objective is the *design of a reliable system for segmenting vocal tract articulators in RT-MRI.*

Articulatory Modeling

The central thesis goal is *the synthesis of vocal tract shape conditioned on the sequence of phonemes to be articulated.* The system design is challenging due to the complex dynamics of the human vocal tract, the large variability in speech articulations, and the physical and acoustical constraints that should be considered.

Model Evaluation

Model evaluation is crucial for any machine learning task. It is necessary to measure the proper dimensions of the problem and evaluating the wrong metric will lead to a waste of resources and problematic models. Also it is important to have the right incentive structures. Machine learning models will fit whatever objective we design, and deep neural networks will search for the "easiest" path towards the minimum. The wrong learning objective will lead to many consequences that we cannot anticipate.

Our work conditions the models in a phonetic sequence; hence, it is natural to expected the model to retain the most phonetic information from the data. Hence, the last thesis objective is *quantifying the phonetic information retained by the articulatory features and reproduced by the vocal tract shape synthesizer.*

4 List of Publications

Publications in international conferences with proceedings

- Vinicius Ribeiro, Yves Laprie. **Autoencoder-Based Tongue Shape Estimation During Continuous Speech.** In *Interspeech*, 2022, Incheon, South Korea [17].

- Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Pierre-André Vuissoz, Yves Laprie. **Towards the Prediction of the Vocal Tract Shape from the Sequence of Phonemes to be Articulated.** In *Interspeech*, 2021, Brno, Czech Republic [18].

Publications in international journals

- Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Jacques Felblinger, Pierre-André Vuissoz, Yves Laprie. **Automatic Segmentation of Vocal Tract Articulators in Real-Time Magnetic Resonance Imaging.** In *Computer Methods and Programs in Biomedicine*, 2023 [19].

- Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Pierre-André Vuissoz, Yves Laprie. **Automatic Generation of the Complete Vocal Tract Shape from the Sequence of Phonemes to be Articulated.** In *Speech Communication*, 2022 [20].

Publications not presented in this thesis

- Yves Laprie, Vinicius Ribeiro, Karyna Isaieva, Pierre-André Vuissoz, Justine Leclere. **Modeling the Temporal Evolution of the Vocal Tract Shape with Deep Learning.** In *ICPhS*, 2023, Prague, Czech Republic [21].

- Vinicius Ribeiro, Yiteng Huang, Yuan Shangguan, Zhaojun Yang, Li Wan, Ming Sun. **Handling the Alignment for Wake Word Detection: A Comparison Between Alignment-Based, Alignment-Free and Hybrid Approaches.** In *Interspeech*, 2023, Dublin, Ireland [22].

- Romain Karpinski, Vinicius Ribeiro, Yves Laprie. **Accelarating the Centerline Processing of Vocal Tract Shapes for Articulatory Synthesis.** In *ICA*, 2022, Gyeongju, South Korea [23].

5 How to read this text

The first three chapters review the state of the art. Chapter 2 discusses the functions of the vocal tract articulators considered in this work, the available data acquisition modalities, their respective advantages and disadvantages, and the public articulatory speech datasets available in the literature. Chapter 3 presents the deep learning methods required to comprehend the thesis, covering general-purpose approaches, image processing, and sequence learning methods. Chapter 4 introduces articulatory synthesis of speech and its variations. We describe its theoretical background and the recent advances using machine learning.

The following two chapters cover the main contributions of the thesis. Chapter 5 presents our approach to tracking the vocal tract articulators during speech from RT-MRI. We explore the usage of a deep convolutional neural network designed for image segmentation to recognize the edges of each articulator under study, then we employ rule-based algorithms to post-process the network's outputs and obtain the precise geometry of the vocal tract.

Chapter 6 presents our approach to modeling speech articulations. We present the two approaches that were developed; the first directly maps the phonetic inputs to the vocal tract shape without the aid of an articulatory model (model-free synthesis), and the second learns a similar mapping but mediated by an autoencoder-based articulatory model. We show that the autoencoder-based model can incorporate constraints on target achievement for consonantal place of articulation.

Moreover, Chapter 6 discusses our model evaluation approach based on phoneme recognition. We train a phoneme recognizer with acoustic and articulatory features to contrast the phonetic information retained by the mid-sagittal contours extracted from the RT-MRI and that recreated by our vocal tract shape synthesizers.

Chapter 7 concludes this thesis by summarizing our main contributions and presenting the directions for future research.

Chapter 2

Articulatory System and Data Acquisition

> The human voice is the most beautiful instrument of all, but it is the most difficult to play.
>
> *Richard Strauss*

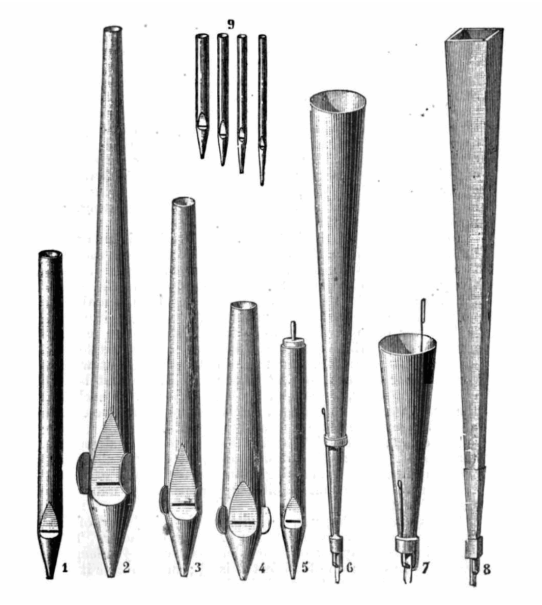
1 Overview

The vocal tract is similar to a wind instrument in that both produce sound by vibrating the air inside a resonator. In a wind instrument, the air is excited by a reed or the musician's lips. The air column length and the resonator's shape determine the instrument's resonance frequencies; therefore, by lengthening or shortening the tube and adding or removing resonating cavities, the musician can play different notes. When pressurized air enters the resonator, it travels at sound speed and is reflected by the instrument's walls. The reflections continue until it forms a standing wave inside the chamber.

A classic illustrative example is the pipe organ (Figure 2.1a). The organ is a wind instrument that produces sound by driving pressurized air through its organ pipes, illustrated in Figure 2.1b. The tubes are tuned to produce different notes, and there are two main types of pipes: flue and reed.

On the one hand, the pitch in flue pipes is controlled by its length and whether the column is open at the end. For an open pipe, the wavelength of the resonating sound is four times the tube's length. Closed pipes produce a sound with a wavelength of twice the tube's length. Hence, a closed tube will resonate at twice the frequency of an open pipe with equal length. On

(a) Organ at the Saint Germain l'Auxerrois Church, Paris. Reproduced from Wikipedia Commons [24]



(b) Drawings of different types of organ pipes. Reproduced from Stone [25]

Figure 2.1: Organ pipes.

the other hand, on reed pipes, the reed's length determines the sound's pitch. In some cases, the reed's length is adjustable, allowing fine-tuning of the instrument.

The human vocal tract also has a source of pressure (the lungs), a reed (the vocal folds), a resonator (the vocal tract itself), and a radiator (the mouth), and voice production follows a similar physical process of pipe organs. However, the human vocal tract is more versatile than a wind instrument in that the shape of the resonator can be actively controlled by the speaker, allowing for more complex variations of sounds that compose human language [26].

The air pressure coming from the lungs towards the vocal tract vibrates the vocal folds at a fundamental frequency determined by their length and tension. The vocal tract shape, determined by many articulators such as the jaw, tongue, and lips, changes over time to create different resonance frequencies. These resonance frequencies, along with the fundamental frequency, determine the voice's pitch, timbre, and loudness. Constrictions between articulators generate plosives and fricative noise, and the nasal cavity, controlled by the velum, adds an extra resonating cavity to give nasality to some phonemes. Then, the mouth radiates the wave sounds into the environment [27].

This Chapter overviews the vocal tract structure from a functional point of view. We briefly describe its anatomy and the structures concerned with this thesis; however, we are not exhaustive and do not intend to discuss biological aspects in detail. Next, we discuss data collection methods for articulatory research. We overview flesh point methods and medical

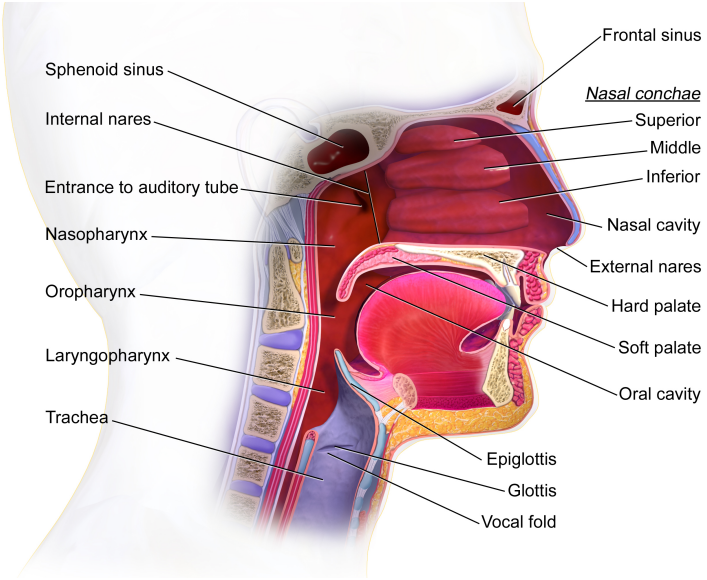


Figure 2.2: The anatomy of the vocal tract and pharynx. Reproduced from Wikipedia Commons [28]

.

imaging approaches, discussing their pros and cons to justify our choice of data modality. Finally, we present the most popular articulatory datasets in the literature, which are relevant and comparable to our research.

2 Vocal Tract Articulations

Figure 2.2 illustrates the vocal tract and pharynx anatomy. Voice production starts as a coordinated action among the diaphragm, abdominal and chest muscles, and rib cage, driving the air from the lungs through the trachea toward the larynx. The larynx is composed of the vocal folds and the muscles and cartilages supporting them. The vocal folds extend from the thyroid cartilage, in the front of the larynx, to the arytenoid cartilage, in the back, which controls the glottis opening and closing. Above the glottis, the epiglottis, a sizeable cartilage connected to the tongue root and the thyroid cartilage, acts like a valve that closes during swallowing to prevent anything from air from entering the lungs, directing food and liquids to the esophagus.

The abduction of the arytenoid cartilage adducts the vocal folds, closing the glottis. The increased pressure of the incoming air opens it, and the air flow vibrates the vocal folds, modulating the sound waves that propagate in the vocal tract. The manipulation of the larynx generates a source sound with a fundamental frequency (`f0`), often referred to as pitch – the perceptive counterpart of `f0`. The pitch can be controlled by altering the tension and the length of the vocal folds. The subglottal pressure controls the voice amplitude (or volume) – higher pressure produces stronger sounds.

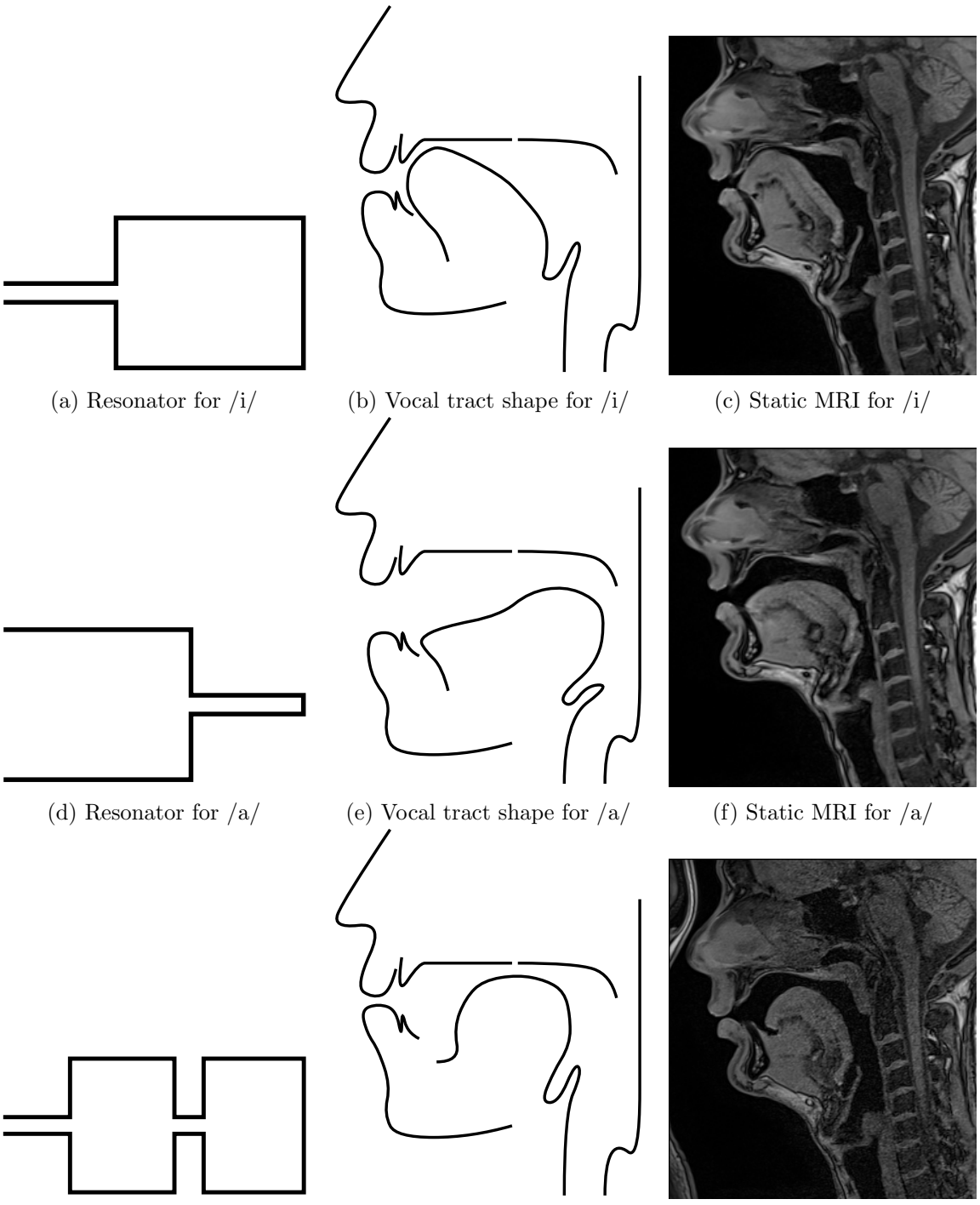
The sound waves propagate through the vocal tract, limited in the back by the pharynx, a portion of tissue that forms the throat. The pharynx extends from the arytenoid cartilage to the nasal cavity. It has minimal movements, essentially vertical displacements, in coordination with the laringeal articulators. The velum controls the airflow through the nasal cavities. The velum is located at the extremity of the soft palate, a soft tissue in the upper back of the mouth cavity connected posteriorly to the hard palate.

When the soft palate contracts, the velum closes the nasopharyngeal port, restricting the sound wave propagation to the oral cavity. During the articulation of nasal phonemes, e.g., /m, n, ã, õ, ɛ̃, œ̃/, the soft palate relaxes, adding an extra resonator to the vocal tract, causing the nasality. Initially, velum lowering was regarded as an independent phonological unit. In contrast, velum raising was seen as a consequence of vocal tract movements during the utterance of oral phonemes. Nevertheless, Blaylock et al. [29] contrasted the velum movements during nasal and oral sounds, showing that temporal coordination of the velum raising during oral stops resembles that of velum lowering during nasal phonemes, suggesting that velum control occurs in either lowering and raising.

The frontal region of the vocal tract is delimited by the hard palate, to which the upper teeth are connected, and the tongue, which moves in coordination with the mandible. The mandible is the largest and strongest bone in the human facial skeleton and the only movable one (discounting the ossicles in the middle ear). It is a primary speech articulator and is connected to the skull by temporomandibular joints, capable of open, protrusion, and lateral movements. Jaw opening, together with tongue movements, performs most of the reshaping of the vocal tract, leading to most of the phonetic contrasts.

The tongue is the largest speech articulator and the one with the most degrees of freedom, ultimately crucial for speech production. The constriction between the tongue dorsum and the hard palate defines palatal consonants, while the constriction between the tongue tip and the alveolar region (back of the upper incisor) defines dental phonemes. Finally, the upper and lower lips define the end of the vocal tract; from there, the energy radiates to the environment. The lips are also responsible for articulating labial consonants.

Figure 2.3 depicts the resonator cavities and the equivalent vocal tract shapes for the vowels /i, a, u/. The coordination of the different articulators modulates speech, changing the fundamental and the formant frequencies and adding nasality to voice. Since multiple strategies can have similar effects on speech, one or more articulators can act to compensate for the absence of movement in another. For example, the jaw strength reduces with aging [30]; thus the tongue compensates for the weaker jaw [31].



(a) Resonator for /i/ (b) Vocal tract shape for /i/ (c) Static MRI for /i/

(d) Resonator for /a/ (e) Vocal tract shape for /a/ (f) Static MRI for /a/

(g) Resonator for /u/ (h) Vocal tract shape for /u/ (i) Static MRI for /u/

Figure 2.3: Schematic representation of the resonating cavities in the left, respective vocal tract shapes in the center, and start MRI of a real speaker for the vowels /i, a, u/. Diagrams inspired by Boë et al. [32]. Static MRI from Douros et al. [33].

This Section briefly described the vocal tract articulators and their roles in speech production. In summary, the articulators that we are mainly concerned about are:

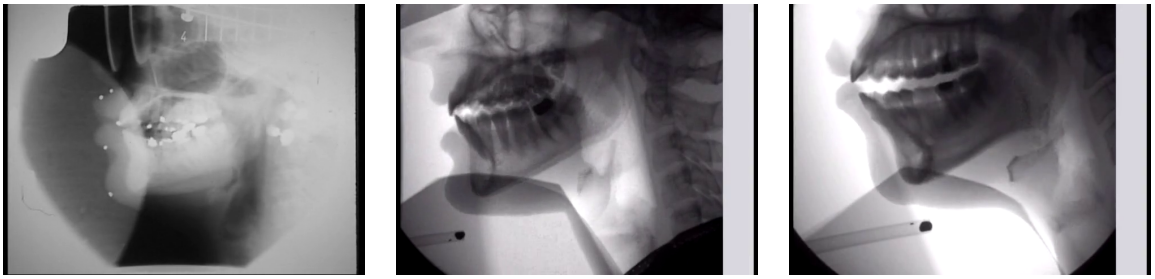
- **Upper and lower lips:** define the radiator;
- **Upper and lower incisors:** delimit the hard palate and the mandible positions;
- **Tongue:** largest vocal tract articulator;
- **Soft palate:** controls the nasal cavity;
- **Pharynx:** delimits the throat;
- **Vocal folds:** defines the source position for voiced phonemes;
- **Arytenoid cartilage, epiglottis, and thyroid cartilage:** delimit the larynx and support the vocal folds.

For a complete anatomical description of the vocal tract, including a comprehensive understanding of the head and neck anatomy, we refer the reader to Hiatt [34].

3 Data Collection Modalities for Articulatory Synthesis

Studying the speech production mechanisms and quantifying articulatory movements require invasive or non-invasive techniques due to most of vocal tract dynamics being not visible from the exterior. The challenges come to finding an approach that allows the collection of large datasets with many hours of speech without altering the articulations, degrading the acoustic signal or harming the subjects' health. An ideal data collection method for articulatory synthesis of speech should have the following characteristics:

- **Coverage:** It should cover the complete vocal tract extension from the glottis to the lips;
- **Time resolution:** It should have a sufficient time resolution for capturing the vocal tract dynamics, including fast constrictions between articulators;
- **Harmless:** It should not present harm or health hazards to the subjects;
- **Naturalness:** It should not perturb the speech, allowing the natural articulation of the phonemes;
- **Non-degradation:** It should not degrade the acoustic or articulatory signals collected;
- **Portability:** It should be portable to work outside the laboratory.



(a) Frame extracted from Brock and Tassi [36]

(b) Frame extracted from Brock and Tassi [37]

(c) Frame extracted from Brock and Tassi [38]

Figure 2.4: Samples extracted from X-ray videos from the DOCVACIM corpus [39].

The existing approaches in the literature range from flesh point tracking to medical imaging. Their application depends on the research field and which articulators are covered. Medical imaging is usually preferred in contexts that require a complete vocal tract visualization, while for silent speech interfaces [35], for example, characteristics such as portability might be more relevant.

3.1 X-Ray Cineradiography

X-ray technology was the most relevant articulatory speech data source since it started being used in the 1920s – see Moll [40] for an early research review. X-ray penetrates the body tissues forming an image in different shades of grey. The tonality is given by the level of radiation absorbed by each tissue. Bones, formed by calcium, absorb the most radiation; therefore, they appear in white in the image, while soft tissues absorb less and appear in grey. Air does not absorb radiation, showing in black.

The drawback of X-ray for speech production studies is related to the low contrast between speech organs, formed mainly by soft tissues, e.g., the tongue and the velum, and to the occlusion of some articulators, as observed in Figure 2.4. X-ray projects the complete vocal tract into a 2D plane, and rigid bodies, such as the jaw, teeth, and dental fillings, might occlude parts of the tongue. Nevertheless, X-rays were abandoned in the 1990s due to the ionizing radiation, which is harmful to the subject, even if the current technology results in much lower levels of absorbed radiation. Before that, large quantities of X-ray films were collected from the 1950s to the 1980s.

Alternatively, X-ray microbeam reduces the exposure to radiation by emitting very narrow and localized X-ray beams to track the movements of tiny pellets attached to specific points in the vocal tract [41]. The exposed area is reduced to around 1 cm^2 per frame, up to 100 frames per second, limiting the radiation absorption. Additionally, the data processing is simplified, even though the complete vocal tract shape is not available, and it requires reconstruction

through interpolation.

3.2 Ultrasonography

Medical ultrasound, or ultrasonography, is a non-invasive medical imaging technique that uses sound waves above the human hearing spectrum (20 to 20 000 Hz). Ultrasonography for clinical usage typically ranges from 2 MHz to 12 MHz [42] – higher frequencies are possible but with a limited ultrasonic penetration depth [43]. The ultrasonography transducer contains piezoelectrical crystals that rapidly vibrate when excited by an electric current. These vibrations emit a sound pulse into the subject's body, which echoes in the human tissues and returns to the probe with tissue-specific reflection properties. When the reflected signals return to the probe, they are reabsorbed by the crystals, emitting electric current, which is used to reconstruct the image [44].

Among many other medical applications, this modality has been used to study tongue gestures [45, 46, 47]. The data is collected by positioning the transducer below the patient's jaw and recording the tongue articulation. For the method's efficiency, the probe's positioning is crucial. For this, two methods exist: immobile and mobile transducers [48]. When the transducer is immobile, its position is fixed relative to the head. The measured movement is the combined displacement of the tongue with the jaw; therefore, the jaw position has to be captured independently and subtracted from the ultrasound measurement.

Nevertheless, the tongue and the jaw are not uniformly coupled along their extension. The angular jaw movement affects the tongue front more than the back; thus, the measurement correction will inevitably be inaccurate. The alternative uses a movable probe that moves with the mandible, and the tongue measurements are relative to the jaw instead of the skull [48].

Ultrasound is advantageous for speech research by registering videos up to 100 Hz, being portable, cheap, and safe. It has been used in speech production studies focused on two-dimensional cross-sectional tongue movements [49, 50, 51] and have several applications for the development of silent speech interfaces [52, 53] However, ultrasonography is restricted to the tongue, missing influential articulations such as the lips, velum, and laringeal articulators, which limits its usage in many areas of articulatory speech research.

3.3 Electromagnetic Articulography

Electromagnetic articulography (EMA) is one of the most extensively used technologies to quantify speech articulatory movements. EMA uses three transmitter magnetic coils to induce a magnetic field around the subject's head and measure the position of a few sensors attached to

specific vocal tract locations. The magnetic field induces a small current in these sensors; since the induced current is inversely proportional to the cube of the sensor's distance to the magnetic coils, the spatial positions of the sensors can be precisely determined. MIT system articulography [54], Movetrack system [55], and Aurora system by Northern Digital (NDI) [56] were the first commercial articulographs. Nowadays, Carstens Medizinelektronik GmbH[1], Bovenden, Germany, is the principal manufacturer of these devices.

Even though the equipment is expensive, the operational cost of EMA is lower than some medical imaging approaches, enabling the acquisition of large multi-speaker datasets. The high spatial precision (0.3 mm for Carstens AG501) and temporal resolution of EMA (200 Hz for Carstens AG500 and 1250 Hz for Carstens AG501), together with the possibility of measuring multiple articulators simultaneously explain its success in speech research. However, it is necessary to point out a few disadvantages. The positioning of the sensors is limited to a few articulators in the oral cavity, excluding pharynx and laringeal articulators, and it is difficult to repeat the same sensor positioning in every acquisition. As Figure 2.5a illustrates, a usual acquisition considers the tongue, lips, mandible, and velum, the latter causing discomfort during the procedure.

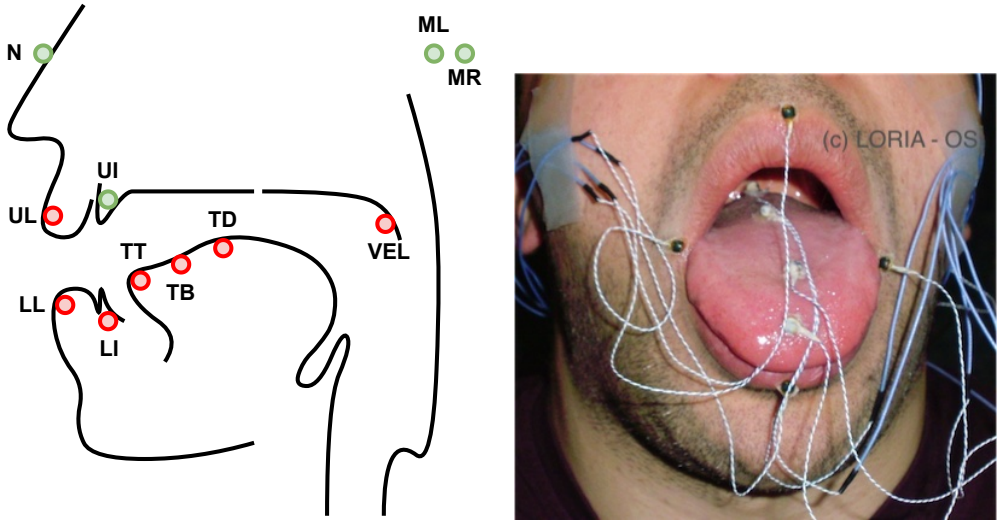
In addition, even if Dromey et al. [57] indicates that speakers adapt in around ten minutes, EMA causes some level of speech impairment due to the wiring coming out of the speaker's mouth, as seen in Figure 2.5b. Moreover, the sensors cannot be too close to each other without interference – Carstens AG500[2] user manual specifies a minimal distance between sensors of 8 mm – limiting the tracking of the entire tongue profile, which is possible with medical imaging.

Concerning safety, EMA is generally a harmless technology and is considered non-invasive. However, some contraindications require attention, as described in the Carstens AG500 user manual. It is not recommended to experiment with test subjects wearing medical appliances such as pacemakers [59] and cochlear implants [60]. Moreover, EMA is not recommended for patients with electromagnetic hypersensitivity, claustrophobia, and immunocompromised patients due to the risk of infection. For a more comprehensive review of EMA practices and procedures, we refer the reader to Rebernik et al. [58].

3.4 Real-Time Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a non-invasive medical imaging technique that captures detailed images of body structures using a large magnet and radio waves. Unlike X-ray, the

[1]Carstens Medizinelektronik GmbH: https://www.articulograph.de/
[2]Carstens AG500 User Manual: http://www.ag500.de/manual/ag500/AG500_manual.pdf



(a) Typical EMA sensors positioning. (b) Speaker in the EMA machine.

Figure 2.5: In (a), the red dots represent the EMA sensors, while the green dots represent reference sensors. The image (a) was inspired by Rebernik et al. [58]. The image (b) was reproduced with authorization from the multimodal platform of the Multispeech team at Loria within the Creativ'Lab.

MRI does not use ionizing radiation and is generally safe for the subject, allowing acquisition sessions of longer than one hour.

The MRI system comprises a primary magnet, responsible for generating an intense and stable magnetic field, and three smaller and weaker gradient magnets to create a variable magnetic field, enabling the scanning of different body parts. In addition, it contains a hydrogen-specific radio frequency (RF) coil. When the patient enters the machine, the primary magnet induces a magnetic field that interacts with the hydrogen atoms in the body, aligning their spins in a particular direction. Next, an RF pulse in a tissue-specific frequency, known as Larmor frequency, is directed toward the region of interest. The protons absorb this energy and spin in a different direction. Then, the gradient magnets are turned on and off to alter the magnetic field at a local level. This method enables the collection of image slices in any direction without requiring the subject to change their position – a great advantage to other image modalities, usually restricted to a single plane. Most often, the acquisitions are made in the transverse (axial), coronal (frontal), and sagittal planes (Figure 2.6). The system fills the Fourier space that is then sampled to reconstruct the image [61].

The machine produces a characteristic noise due to the opposite direction between the main magnetic field and the electric current going through the gradient magnets' wires, which is problematic for acquiring speech databases. Two optical microphones are necessary for the recording; one is positioned close to the subject's mouth, while the second is placed farther away to record the environmental noise. Then, a denoising algorithm filters out the MRI noise, with

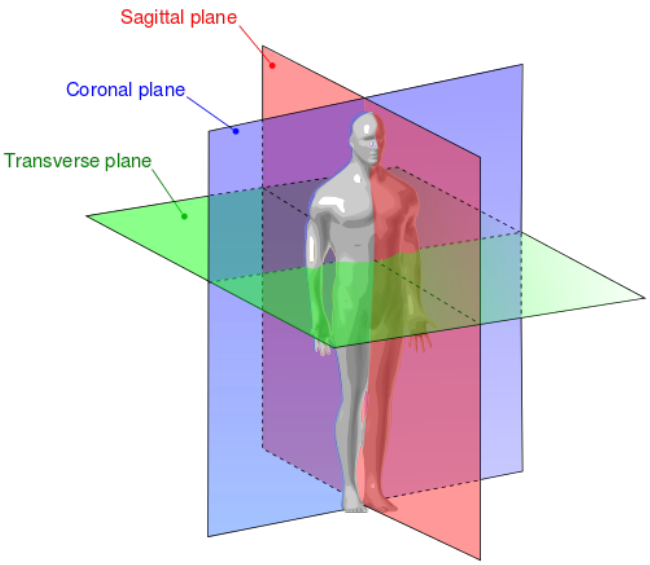


Figure 2.6: Anatomical planes for medical imaging. Reproduced from Wikipedia Commons [62].

the disadvantage of attenuating frequencies in the signal, distorting the speech, and harming the final analysis.

Since the first applications to vocal tract measurement [63, 64], the interest in these approaches has grown. It is now the dominating technique in the field. However, the high acquisition time poses challenges to studying speech production. In the initial studies [64], MRI was used to collect vocal tract images for sustained vowels, with an overall acquisition time of 45 seconds, later decreasing to around five seconds with lower image quality.

Over the years, the acquisition time has been reduced, allowing the collection of dynamic processes in the human body such as heartbeat [65] and speech [66]. Real-time MRI (RT-MRI) has been increasingly used in vocal production research since it allows a one-plane visualization of the vocal tract from the glottis to the lips with a sampling rate of 50 fps. However, it still presents a few drawbacks. First, it is substantially more expensive than the other techniques. Second, MRI technology is only sensitive to structures containing water; thus, bones, with short T2*, are indistinguishable from air in the images. Third, the MRI machine is claustrophobic, being prohibitive for some subjects. Fourth, the subjects cannot have any ferromagnetic material in their bodies, including prostheses, dental braces, and pacemakers. Fifth, the supine position and the prolonged sustenation cause hyperarticulation, distortion in the tongue movement and lack of velum control [67].

Additionally, some characteristics of static MRI must be abandoned to achieve a high temporal resolution. Static MRI provides a high-resolution 3D image, but the high acquisition time limits the use of 3D real-time recording, even though attempts can be found in the literature [68, 69]. Thus, for RT-MRI, the acquisition is limited to one plane – typically the

(a) (b) (c)

Figure 2.7: Difficult cases of RT-MRI acquisition, challenging the image processing.

mid-sagittal – and a smaller image resolution – typically 136 × 136 pixels. Still, RT-MRI is state-of-the-art for vocal tract observation and has been widely used for the study of speech, singing [70], blowing wind instruments [71], and human beatboxing [72, 73]. Lingala et al. [74] presents guidelines, technical considerations, and recommendations for RT-MRI for studying speech.

Even though RT-MRI was a significant step forward for speech research, processing and analyzing those images are challenging tasks. The sampling frequency is typically too low for the articulators' velocity during the natural speech, usually causing blurred images (Figure 2.7a). Also, when two articulators are in contact, it might be hard to differentiate between them (Figure 2.7b). Moreover, the pixel spacing and the slice thickness are relatively large, around 1.5 to 2 mm for the first and around 8 mm for the latter. The slice thickness causes a problem known as partial volume effect [75], when the entire slice volume is projected into the same plane, producing an uncertainty relative to the actual articulator position in the mid-sagittal plane. Finally, the laringeal articulators are hard to analyze due to the narrow area and their fast movements (Figure 2.7c).

3.5 Summary

Table 2.1 consolidates the characteristics of each data acquisition modality presented in this Section. As it was exposed, none of the available modalities checks all the desired boxes. Medical imaging provides the most comprehensive approaches; however, drawbacks must be considered in developing speech corpora. Due to ionizing radiation, X-ray is not used anymore. Ultrasound has limitations regarding the field of view, and RT-MRI is the most expensive and least portable method. Nevertheless, the presented benchmark helps guide the choice of data modality for research.

Table 2.1: Description of the characteristics present for each data modality. [1] The supine position cause a few articulatory distortion, harming the naturalness. [2] The positioning of the ultrasonic probe reduces the jaw opening.

	X-ray Cineradiography	Ultrasound	EMA	RT-MRI
Coverage	Complete	Limited	Limited	Complete
Time resolution	50 Hz	100 Hz	1250 Hz	50 Hz
Harmless	No	Yes	Yes	Yes
Naturalness	Yes	Yes	No	Yes [1]
Non-degradation	Yes	Yes [2]	No	No
Portability	No	Yes	No	No

4 Articulatory Speech Datasets

X-ray was one of the earliest articulatory data collection modalities, with large quantities of films being collected between the 1950s and 1980s. Nevertheless, these films deteriorate with time and need proper storage conditions. To preserve these data, Munhall et al. [76] compiled a dataset of X-ray films for speech research containing 55 minutes of speech from 14 subjects (seven males and seven females) in Canadian English and French. A few years later, Westbury et al. [77] released the University of Winsconsin X-ray Microbeam Database (XRMB) with the speech of 47 speakers in American English, with about 20 minutes of recordings per speaker. Even after decades of publication, the XRMB database is still extensively used [78, 79] due to the large speaker variability. Data availability is often insufficient, requiring software and standardized procedures to process it. In this regard, Sock et al. [80] released the DOCVACIM X-ray database and the tools and procedures to exploit it for studying speech production.

After the abandonment of X-rays, EMA rapidly grew in popularity. The MOCHA-TIMIT articulatory database [81] was one of the earliest releases, containing 460 utterances in British English for two speakers – one male and one female. Additionally to the EMA data, the corpus includes frontal videos of the mouth region. Later, the `mngu0` database [82] introduced a more extended set of EMA recordings, providing 1 354 utterances in British English for a single-speaker, totaling 67 minutes of speech recorded at 200 Hz. The availability of these sets was an essential step towards a better understanding of speech production.

The growing popularity of real-time MRI encouraged the release of the MRI subset of the `mngu0` database [83] containing volumetric MRI of sustained phonemes and RT-MRI movies with repetitions of consonant-vowel (CV) syllables. Later, Narayanan et al. [84] published the USC-TIMIT, an extensive RT-MRI database with ten speakers in American English, which was later extended to the USC-EMO-MRI, an RT-MRI database of emotional speech.

The USC-EMO-MRI database contains midsagittal films of ten speakers and emotional

labels that include neutral, anger, happiness, and sadness. The speakers were asked to immerse themselves in one of the target emotions and read the "Grandfather" passage plus seven shorter sentences in English. The emotion quality of the speech was evaluated by ten actors and actresses that, after listening to the passage, provided their opinion regarding which emotion best represents the passage, their confidence in that opinion, and an emotional strength level to the passage. Emotional speech is a growing research topic and the release of USC-EMO-MRI permits it to be studied in the articulatory level, being an important advancement in the field.

Later, Sorensen et al. [85] released a database containing the speech of 17 speakers in American English. The dataset includes the RT-MRI movies with the recorded denoised audio plus volumetric MRI of sustained phonemes. The available utterances includes repetitions of consonant-vowel-consonant (CVC) and vowel-consonant-vowel (VCV) utterances, read passages and spontaneous speech.

Most of the available data in speech sciences are in English – usually American English, which limits the speech research in other languages. In this regard, Teixeira et al. [86] released an RT-MRI dataset for European Portuguese with one female native speaker. The dataset was mainly designed to characterize nasal vowels in a wide range of phonetic contexts. Douros et al. [33] released a database in French containing the data of two native French speakers. The database contains static MRI of sustained phonemes as well as RT-MRI of complete sentences plus VCV repetitions. Later, Isaieva et al. [87] published a larger set containing the data of ten native French speakers, including read sentences and VCV repetitions. The methods from Douros et al. [33] and Isaieva et al. [87] are particularly relevant for this thesis since they were largely explored in its related publications.

Most recently, Lim et al. [88] released the most extensive dataset so far, containing 75 American English speakers, with an average of 17 minutes of speech per subject. Such large dataset corresponds to a significant advancement in articulatory speech research due to the possibility of studying inter-speaker variability in a much wider range. Table 2.2 summarizes the characteristics of the most popular articulatory speech datasets in the literature.

5 Conclusion

This Chapter briefly covered the anatomy of the vocal tract, presenting the main articulators that concern our work and discussing their functions in speech production. Moreover, we reviewed the most relevant data collection modalities to study vocal tract dynamics. We discussed their advantages concerning a list of characteristics desired in an ideal system. The literature review shows that despite no method so far concentrating all of them, the RT-MRI is the pre-

Table 2.2: Summary of most popular articulatory speech databases.

	Dataset	Modality	Language	Number of Speakers
XRMB	Westbury et al. [77]	X-ray Microbeam	English (US)	47
DOCVACIM	Sock et al. [39]	X-ray	French	
MOCHA-TIMIT	Wrench [81]	EMA	English (UK)	2
`mgnu0`-EMA	Richmond et al. [82]	EMA	English (UK)	1
`mgnu0`-MRI	Steiner et al. [83]	RT-MRI	English (UK)	1
	Teixeira et al. [86]	RT-MRI	Portuguese (Europe)	1
USC-TIMIT	Narayanan et al. [84]	RT-MRI	English (US)	10
USC-EMO-MRI	Kim et al. [89]	RT-MRI	English (US)	10
	Sorensen et al. [85]	RT-MRI	English (US)	17
	Douros et al. [33]	RT-MRI	French	2
	Isaieva et al. [87]	RT-MRI	French	10
	Lim et al. [88]	RT-MRI	English (US)	75

ferred choice in research due to the complete coverage of the vocal tract, the satisfactory time resolution, and being harmless to the subject. The challenges presented by the choice of RT-MRI were detailed, with illustrative examples. Finally, we presented a review of the available datasets in the literature. We discussed the characteristics of each corpus regarding modality, language, and number of speakers. We have also briefly detailed their contributions to articulatory speech research.

Chapter 3

Deep Learning

> All models are wrong, but some are useful.
>
> *George Box*

1 Overview

Deep learning is the epicenter of the current computer science and artificial intelligence (AI) research. Large Language Models (LLMs) such as BERT [90], GPT-3 [91], and LLaMA [92], which are capable of reproducing language at a human level, are causing extensive discussions in the academy and industry around the capacities of AI. Generative visual models such as DALL-E 2 [93] shocked the community by generating photorealistic images from human prompts. While taking giant steps forwards, the advancement of deep learning research raises debates around ethics, fairness, AI's ecological footprint, human rights, intellectual property, and others. Even though the rise in attention towards these models is recent, neural networks have existed for many years.

In the beginning, the brain's functioning inspired the development of neural networks. Its primary structures were first thought of in 1943 by McCulloch and Pitts [94], which were trying to copy the behavior of the human neuron. Later developments led to the creation of the perceptron [95], whose limitations meant a first drawback in the history of artificial neural networks. The simple structure of the perceptron seemed insufficient to model elementary boolean operations. Nevertheless, it would later show its usefulness by combining other perceptrons in layers to build more extensive networks, giving birth to the multilayer perceptron.

The field solidified by developing a robust learning method known as *backpropagation algorithm* [96]. The gradient of the prediction error w.r.t. the network's weights is calculated using

the chain rule and used as an update rule. Nevertheless, further challenges were still appearing. The problem of vanishing gradients, when the gradients become smaller at every layer that the weights stop changing, or change very slowly, was prohibitive for deeper networks with more than a few hidden layers. In addition, deep neural networks were regarded as exceedingly data-consuming and computationally inefficient. These difficulties seemed insurmountable, which caused neural network research to cool down for decades.

After a long time of disinterest by the scientific community, the AI winter, in 1998, Yann LeCun proposed the LeNet [97], one of the earliest convolutional neural network architectures for solving the handwritten digit recognition task, on the `MNIST` dataset [98]. However, many scientists still left deep neural networks aside due to the number of resources required for training. It was only by 2012 when the AlexNet [99], proposed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, conquered the ImageNet Large Scale Visual Recognition Challenge [100] by a 10.8% margin compared to the second place, that the community's eyes turned directly to deep learning. With larger datasets and recent advances in graphical processing units (GPU), which performs matrix operations more efficiently than traditional CPUs, the ingredients for a deep learning revolution were available.

This chapter will review the leading deep learning methods and architectures. The review focuses on the methods concerning this thesis's scope. Nevertheless, we may extend the discussion to a few topics that were not employed in our experiments for completeness. For a more comprehensive understanding of traditional deep learning methods, we refer the reader to Goodfellow et al. [101].

2 General Purpose Deep Learning Methods

2.1 The Perceptron

The artificial neural networks that are common nowadays began as simple computational models of the brain, with two models of cognition: associationism – the idea that mental processes operate by association between one mental process to its successor – and connectionism – the idea that neurons connect to other neurons and the connection strength changes according to past experiences; the second being a more successful cognition model.

In human biology, the neuron is an electrically excitable cell that transmits electric signals across a network through a process called *synapse*. In a simplified form, the neuron comprises the dendrites, the soma, and the axon. When the axon terminals of a source neuron connect to the dendrites of a target neuron, an electric pulse flows through the target neuron, propagating

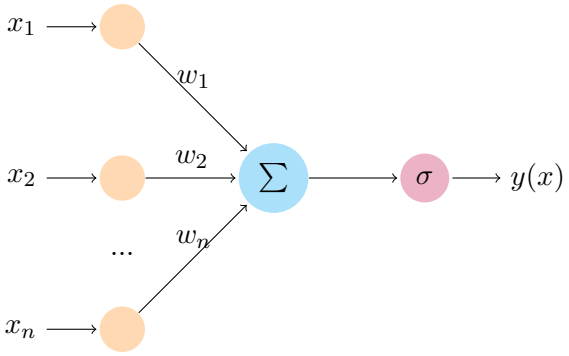


Figure 3.1: Perceptron model proposed by Rosenblatt [95]. In the model, σ represents a thresholding function. Reproduced from Raj [103].

the signal along the chain. The biological neuron inspired its first computational neuronal model, created by McCulloch and Pitts [94]. Although innovative, the idea lacked a learning mechanism essential to these algorithms. The connectionist model of cognition inspired Hebb [102] to develop the first learning algorithm that follows the premise that "neurons that fire together wire together", formally described by [103]

$$w_{ij} = w_{ij} + \eta x_i y_j$$

where w_{ij} is the weight of the $i^{\texttt{th}}$ input (x_i) to the $j^{\texttt{th}}$ output (y_j). In simplified terms, whenever x_i and y_j have a non-zero value simultaneously, their connection gets updated by a factor of η.

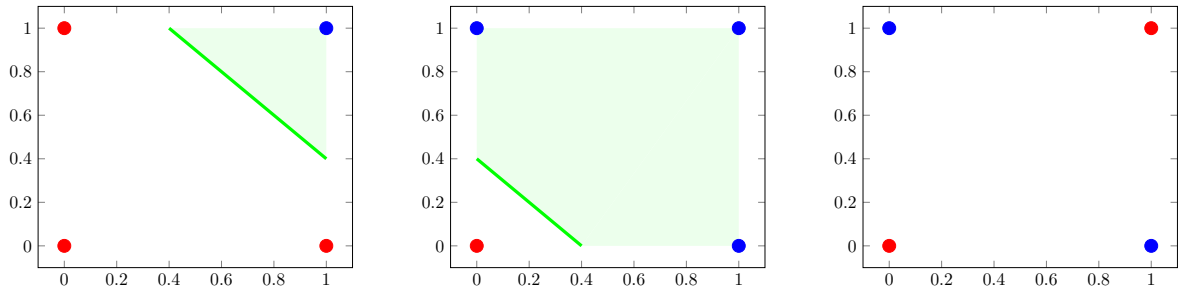
In 1958, Rosenblatt [95] proposed a more advanced model, the perceptron, visually represented by Figure 3.1. The perceptron is a neural model that fires if the combined inputs exceed a threshold, whose mathematical model is given by [103]

$$y(x) = \begin{cases} 1, & \text{if } \sum_{i=1}^{n} w_i x_i + b \geq T \\ 0, & \text{otherwise} \end{cases}$$

where T is the threshold value and b is the bias. A keystone for the success of Rosenblatt [95] was the proposal of a convergent learning mechanism that updates the weights whenever the output is wrong. The learning algorithm is given by [103]

$$w = w + \eta(d(x) - y(x))x$$

where $d(x)$ is the target output and $y(x)$ is the perceptron's response to the input x. Unlike the Hebbian approach, this mechanism is based on prediction errors instead of the co-occurrence of neuronal triggers.



(a) AND operation (b) OR operation (c) XOR operation

Figure 3.2: Linear separability of basic boolean operations.

Table 3.1: Truth tables for AND, OR, and XOR operations.

X	Y	$X \times Y$ ((AND))	$X + Y$ (OR)	$X \bigoplus Y$ (XOR)
0	0	0	0	0
0	1	0	1	1
1	0	0	1	1
1	1	1	1	0

The perceptron was a revolutionary model, capable of replicating boolean operations such as AND and OR (Table 3.1). However, Minsky and Papert [104] demonstrated that the perceptron fails to mimic the XOR operation, configuring a critical setback for the idea. As Figure 3.2a and Figure 3.2b show, the AND and OR operations are linearly separable operations; hence they can be computed by the perceptron. However, as shown by Figure 3.2c, separating the XOR by a line is impossible, meaning that a single neuron is insufficient. Non-linear operations require networked elements, giving birth to the multilayer perceptron.

The multilayer perceptron (MLP) was an evolution of the single perceptron system, consisting of chaining together multiple perceptrons in different layers to form a network. The perceptrons are organized in layers, each containing several of these units. The outputs of the preceding layers are the inputs to the following ones, allowing the computation of arbitrarily complex boolean functions, being provably a *universal boolean approximator*, i.e., any truth table can be expressed in the form of a one-hidden-layer MLP [103]. Figure 3.3 illustrates the MLP architecture proposed by Minsky and Papert [104], which can reproduce the XOR truth table (Table 3.1) with the appropriate set of weights. However, the number of neurons of an XOR network with N input variables will require $2^{N-1} + 1$ perceptrons, growing exponentially with input size. The solution comes in the former of deeper networks. By reorganizing the perceptrons in multiple hidden layers, the number of perceptrons is reduced to $3(N-1)$ arranged in $2\log_2(N)$ layers, growing linearly with the input size [103].

The universal boolean approximator characteristic allows using the MLP as a classifier net-

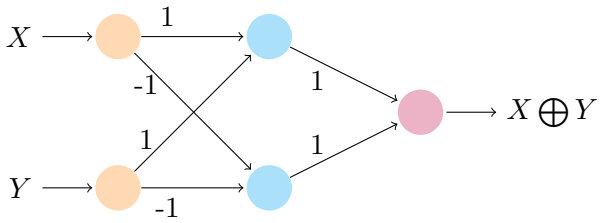


Figure 3.3: Multilayer perceptron for performing `XOR`. Reproduced from Raj [103].

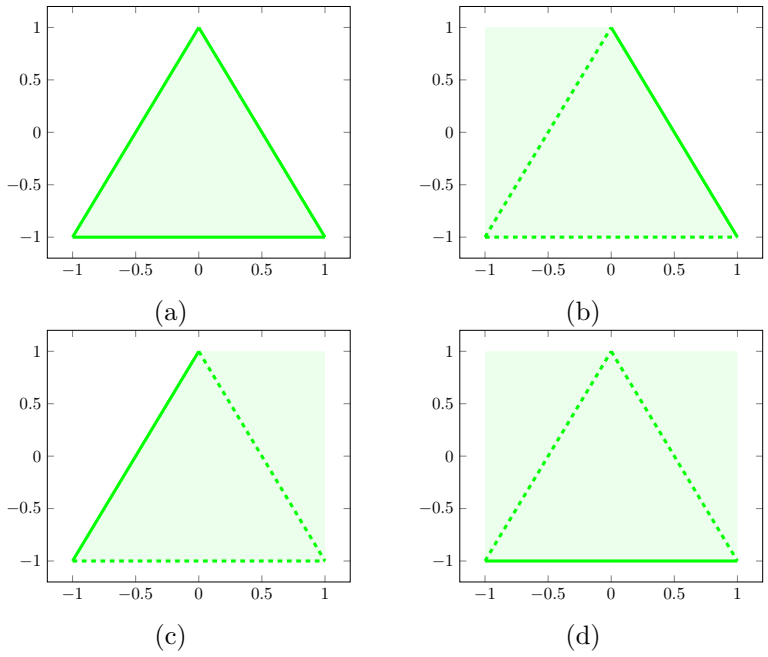


Figure 3.4: Triangular decision boundary. Inspired by Raj [103].

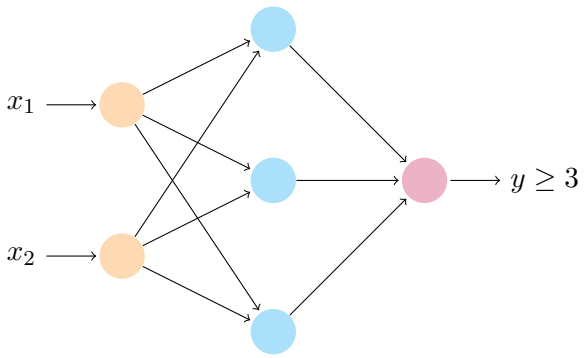


Figure 3.5: MLP architecture for a triangular decision boundary. Inspired by Raj [103].

work, composing arbitrarily complex decision boundaries with arbitrary precision, limited only by the number of perceptrons used to represent the classifying function. A simple exercise to understand it is in the triangular decision boundary from Figure 3.4a. Although the shaded area is not linearly separable, it can be decomposed into three linear decision boundaries (Figure 3.4b, Figure 3.4c, Figure 3.4d), each of them can be approximated by a single perceptron. As seen in Figure 3.5, a succeeding neuron performs the `AND` operation on the outputs of the individual preceding neurons, firing when $y \geq 3$ (for three linear frontiers) and approximating the triangular region. This exercise can be extrapolated for any geometry with additional neurons in the hidden layer or extra hidden layers, making the MLP a *universal classifier* [103]. Likewise the boolean case, deeper MLPs require fewer neurons than a single-layer network. However, the power of the MLP goes far beyond modeling boolean functions and decision boundaries.

Adding continuous activation functions allows mapping the MLP output to interpretable real values. Common activations are the sigmoid, hyperbolic tangent, and the rectified linear unit (`ReLU`). Including activation functions in the models produces a graded neuron output instead of a discrete result. These graded outputs permit information propagation in the network in the form of gradients. The activation functions qualify the MLP as a *universal function approximator* provided that it has *sufficient capacity* [103], i.e., an MLP is incapable of modeling a function that has more convex regions than the capacity of the network, meaning that any try will result in an approximation with non-zero error. The capacity of a neural network is typically measured in terms of its Vapnik-Chervonenkis dimension (VC-dimension) [105].

2.2 Backpropagation

A major advancement in developing artificial neural networks was the backpropagation algorithm – a learning procedure that repeatedly adjusts the weights between the network's connections to minimize a pre-defined error metric between the network's outputs and the desired targets [96]. Backpropagation uses the chain rule to compute the effect of varying the network's weights in the prediction error.

Consider a neural network with L layers that takes an input vector $x \in \mathbb{R}^{M_{\text{in}}}$ and outputs a prediction $\hat{y} \in \mathbb{R}^{M_{\text{out}}}$. Let $z^{(l)}$ be the output of layer l and $a^{(l)}$ the output of the activation function of layer l. $W^{(l)} \in \mathbb{R}^{M_l \times M_{l-1}}$ and $b^{(l)} \in \mathbb{R}^{M_l}$ represents the weights matrix and the bias vector at layer l, respectively, and M_l is the number of neurons at layer l. Then, the backpropagation can be computed with the following steps:

1. Propagate the training input through the network to get the predicted output.

$$
\begin{aligned}
a^{(1)} &= x \\
z^{(l)} &= W^{(l-1)} a^{(l-1)} + b^{(l-1)} \\
a^{(l)} &= \sigma(z^{(l)}) \\
\hat{y} &= W^{(L)} a^{(L)} + b^{(L)}
\end{aligned}
$$

   where σ represents a non-linearity [106].

2. Compute the prediction error between the expected output y and prediction $\hat{y}$.

$$E = \mathcal{L}(y, \hat{y})$$

   where $\mathcal{L}$ is a pre-defined loss (or cost) function.

3. Compute the derivatives of the error w.r.t. the network's weights, which describes how a change in the weights will affect the prediction error.

   For a single weight, the derivative is [106]

$$
\begin{aligned}
\frac{\partial E}{\partial w_{jk}^{(l)}} &= \frac{\partial E}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial w_{jk}^{(l)}} \\
z_j^{(l)} &= \sum_{k=1}^{M_{l-1}} w_{jk}^{(l)} a_k^{l-1} + b_j^{(l)} \\
\frac{\partial z_j^{(l)}}{\partial w_{jk}^{(l)}} &= a_k^{l-1} \\
\frac{\partial E}{\partial w_{jk}^{(l)}} &= a_k^{l-1} \frac{\partial E}{\partial z_j^{(l)}}
\end{aligned}
$$

   Likewise, for a single bias, the derivative is [106]

$$
\begin{aligned}
\frac{\partial E}{\partial b_j^{(l)}} &= \frac{\partial E}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial b_j^{(l)}} \\
\frac{\partial z_j^{(l)}}{\partial b_j^{(l)}} &= 1 \\
\frac{\partial E}{\partial b_j^{(l)}} &= a_k^{l-1} \frac{\partial E}{\partial z_j^{(l)}}
\end{aligned}
$$

4. Update the weight in the opposite direction of the gradients.

$$\Delta w_{jk}^{(l)} = -\epsilon \frac{\partial E}{\partial w_{jk}^{(l)}}$$
$$\Delta b_{j}^{(l)} = -\epsilon \frac{\partial E}{\partial b_{j}^{(l)}}$$

where ϵ is the learning rate.

The procedure is repeated until convergence. The drawback of the backpropagation algorithm is the case when the loss function is not convex. Then, gradient descent is not guaranteed to find the global minimum. Various methods focus on avoiding local minima in neural networks, including optimization strategies such as stochastic gradient descent [107] and learning rate scheduling policies [108].

Current deep learning frameworks do not require the explicit computation of the gradients anymore. Instead, only the network's forward pass and the loss function are explicitly defined; the backward pass is computed using automatic differentiation [109]. Automatic differentiation is a great advancement for the practice of deep learning, enabling complex network architectures and loss functions without requiring defining their derivatives.

2.3 Autoencoder

Learning a compact and meaningful manifold from multi-dimensional data is paramount in machine learning research. However, the training of such models is complex due to the difficulty in finding a task that produces general features and the scarcity of labeled data. Initially, these representations were learned using benchmark datasets, usually designed for classification tasks. However, this approach only fits well in some cases. Most often, a large enough dataset will not exist, or the pre-built representations will not fit properly the target application, and fine-tuning will be required. The alternative becomes learning unsupervised or self-supervised representations, which do not require an annotated dataset. Since collecting data is usually *much easier* than gathering the corresponding labels, it is the preferred approach for most applications.

The autoencoder [110, 111, 112] is a neural network that attempts to copy its inputs to its output [101]. Since the target of the autoencoder is the input data itself, it learns to find a representation that retains most of the data variance in an unsupervised manner. A general autoencoder structure contains an encoding network (g_{enc}), an information bottleneck (the latent space), and a decoder network (g_{dec}) (Figure 3.6). The learning objective of the

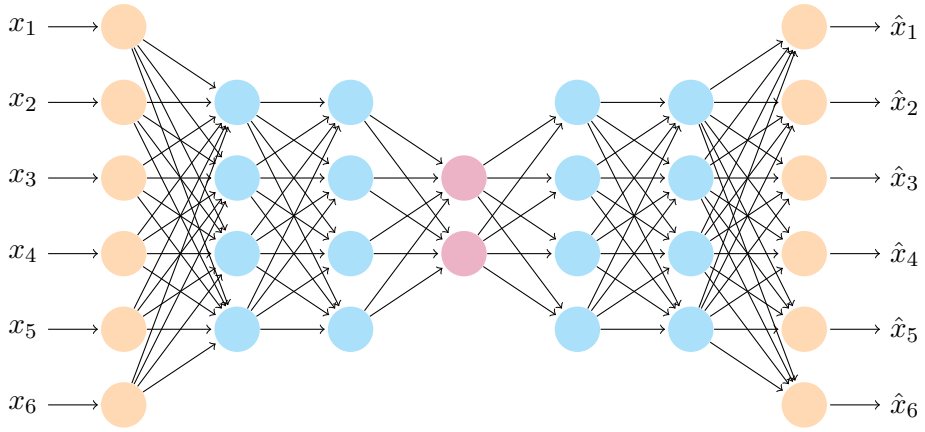


Figure 3.6: General architecture for an autoencoder. The orange nodes represent the input (left) and output (right) layers, the cyan nodes represent to the hidden layers, the pink nodes represent the latent space.

autoencoder becomes

$$\hat{x} = g_{\text{dec}}(g_{\text{enc}}(x))$$

$$arg\,min_{g_{\text{enc}},g_{\text{dec}}} \text{E}[\mathcal{L}(x,\hat{x})]$$

where E is the expected value and $\mathcal{L}$ is a reconstruction error function, usually designed as the L_2-norm [113]. Since the latent space has a much more limited dimensionality than the inputs, the bottleneck encodes only the essential information for the reconstruction performed by the decoder. Therefore, the autoencoder is often seen as a non-linear dimensionality reduction algorithm.

If $g_{\texttt{enc}}$ and $g_{\texttt{dec}}$ are only composed of linear layers, it will result in a linear autoencoder [114], and by removing the non-linear activations, it should learn a similar representation as the one provided by principal components analysis (PCA) [115]. However, unlike PCA, the autoencoder lacks essential characteristics such as orthogonality and statistical independence between the components. Additionally, the autoencoder latent space is not ranked, meaning the variance is distributed among its components in no particular order. Since it does not have the same guarantees as PCA, the autoencoder will use its representational power to encode all data variability simultaneously.

A PCA-like autoencoder was proposed by Ladjal et al. [116] using a specific algorithm for training. The proposed algorithm trains each component individually and freezes the already-trained weights while fitting the following ones. Since the network can only learn one component at a time, it must encode the largest source of variability that has yet to be learned, achieving ranked and independent components.

Autoencoders have many applications in machine learning such as denoising [117, 118], classification [119, 120], clustering [121, 122], and anomaly detection [123].

2.4 Variational Autoencoder

The traditional autoencoder is a deterministic machine, meaning that it does not model the probability distribution of the data; thus, it does not serve as a generative procedure. The latent space will encode the data variability to be easily reconstructed. However, it can only interpolate among the samples in training set. Kingma and Welling [124] significantly improved the model's representation capacity by creating a version based on the variational Bayes inference named variational autoencoder (VAE). The variational autoencoder is a generative model that attempts to learn the probability distribution that governs the data generative process.

Let $X = \{x_i\}_{i=1}^{N}$ be a set of independent and identically distributed observations in the data space. The VAE assumes a generative process given by a set of parameters θ for x conditioned on the latent space z – $p_\theta(x \mid z)$ – referred to as the *likelihood*. The latent vector z is drawn from a *prior* distribution $p_\theta(z)$, and its *true posterior* distribution is denoted as $p_\theta(z \mid x)$. Nevertheless, typically the marginal likelihood $p_\theta(x) = \int p_\theta(z)p_\theta(x \mid z)$ and the true posterior distribution $p_\theta(z \mid x) = \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)}$ are intractable. In these cases, the intractability prohibits the usage of expectation maximization or variational Bayes algorithms. Moreover, when large datasets are available, it is costly to compute the parameters θ.

The VAE proposed by Kingma and Welling [124] presents an efficient algorithm. It proposes a recognition model $q_\phi(z \mid x)$ that approximates the intractable true posterior $p_\theta(z \mid x)$. The parameters θ and ϕ are unknown and are learned together in a joint procedure. In the deep learning literature, the recognition model ($q_\phi(z \mid x)$) is frequently referred to as the *encoder* and the likelihood ($p_\theta(x \mid z)$) is referred to as the *decoder*.

The marginal log-likelihood of the data $p_\theta(x)$ becomes

$$\log p_\theta(x) = D_{KL}(q_\phi(z \mid x) \| p_\theta(x \mid z)) + \mathcal{L}(\theta, \phi; x)$$

where the first term refers to the Kullback-Leibler (KL) divergence between the approximate and the true posterior distributions. Since the KL divergence is non-negative, the second term is the lower-bound for the marginal probability w.r.t. θ and ϕ; therefore, it is referred to as the

variational lower-bound. $\mathcal{L}(\theta, \phi; x)$ can be re-written as

$$\log p_\theta(x) \geq \mathcal{L}(\theta, \phi; x) = \mathrm{E}_{q_\phi(z|x)}[-\log q_\phi(z \mid x) + \log p_\theta(x, z)]$$
$$\mathcal{L}(\theta, \phi; x) = D_{KL}(q_\phi(z \mid x) \| p_\theta(x \mid z)) + \mathrm{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)]$$

Optimizing $\mathcal{L}(\theta, \phi; x)$ w.r.t. θ and ϕ is necessary, but the gradients w.r.t. ϕ can be very unstable and the optimization becomes impractical. To solve this problem, Kingma and Welling [124] introduced the *reparameterization trick*.

The reparameterization trick is an alternative method for sampling from $q_\phi(z \mid x)$. Let $z \sim q_\phi(z \mid x)$ be a continuous random variable. We can express z as a deterministic random variable $z = g_\phi(\epsilon, x)$ where $\epsilon \sim p(\epsilon)$ is an independent auxiliary variable and g_ϕ is a function parameterized by ϕ. The method is appropriate since we can re-write the expectation of $q_\phi(z \mid x)$ such that the Monte Carlo estimation is differentiable w.r.t. ϕ. For the univariate Gaussian case, z is given by $z \sim p(z \mid x) = \mathcal{N}(\mu, \sigma^2)$, and the reparameterization of z becomes $z = \mu + \sigma\epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ [124].

In practice, the true posterior distribution $p_\theta(z \mid x)$ can be approximated by $q_\phi(z \mid x) = \mathcal{N}(h_\mu(x), h_\sigma(x))$, where $h_\mu(x)$ and $h_\sigma(x)$ are the mean and the covariance of the distribution estimated by the encoder network. The sampling from the approximate posterior will be $z = h_\mu(x) + h_\sigma(x)\epsilon$, with $\epsilon \sim \mathcal{N}(0, 1)$ [113]. If we denote the decoder function as g_{dec}, the loss function becomes

$$\mathcal{L}(x, z) = \mathcal{L}_{\texttt{rec}}(x, g_{\text{dec}}(z)) + D_{KL}(\mathcal{N}(h_\mu(x), h_\sigma(x)), \mathcal{N}(0, 1))$$

where $\mathcal{L}_{\texttt{rec}}$ is a reconstruction loss designed as the L_2-norm, and g_{dec}, h_μ, and h_σ are optimized using backpropagation [113].

Applications of variational autoencoders are vast, including classification tasks [125, 126] and audio source separation [127], but are especially useful for generative tasks due to the fast and tractable sampling and the easy access to the encoding network [128].

2.5 Knowledge Transfer in Deep Neural Networks

Data availability is a determinant factor in deep learning. Nevertheless, in the real world, well-annotated and curated datasets are usually unavailable, especially domain-specific data. Thus, techniques to deal with the lack of specialized datasets have proliferated in the literature, and many have become standard practices in deep learning. Among these methods, knowledge transfer is one of the most popular.

Knowledge transfer is based on the intuitive idea that different tasks might require similar abilities. Humans experience it all the time. When learning to ride a motorcycle, we will reuse the balancing skills from when we first learned how to ride a bicycle. When learning a second language, we will invariably make associations with our mother tongue; the more languages we learn, the easier it is to learn the successive ones since we have a broader linguistic skill set. The initial approaches to knowledge transfer were based on the intuition that the earliest features learned by the neural network might be helpful to an entire family of tasks regardless of the final objective. Hence, it would be possible to reuse the initial layers pre-trained with a large general-purpose dataset on training a specialized task without enough available data.

A classic example of knowledge transfer comes from computer vision. Suppose we are challenged to classify a set of animal pictures into three categories: felines, canines, and birds. Among the features that might be useful for the task are the number of legs, presence of wings, presence of a beak, facial shape, and ears shape. These features are particular for the animals that we are considering. They probably would not be present if we transfer knowledge from a dataset such as the Cityscapes [129], which focuses on understanding urban street scenes. Nevertheless, these high-level and task-specific features can be decomposed into mid-level features such as circles, triangles, and rectangles, which in turn can be further decomposed into an even lower level, such as corners, horizontal borders, vertical borders, diagonals, and curvatures. That is precisely how a deep neural network will learn to recognize patterns in the data; it starts from elementary structures, learned by the initial layers, then the further layers compose these basic patterns into more complex ones up to the point that the final layers have learned how to recognize high-level features.

Initial knowledge transfer approaches exploited the idea of re-utilizing the knowledge from a source task in a different but related task by copying the weights learned with the first task to the network of the second; then the training for the second task continues as usual; a process usually referred to as *pre-training with fine-tuning* approach [130]. Even though there are still a substantial amount of specialized weights to fit, the pre-initialized network massively reduces the number of parameters, improving performance and generalization.

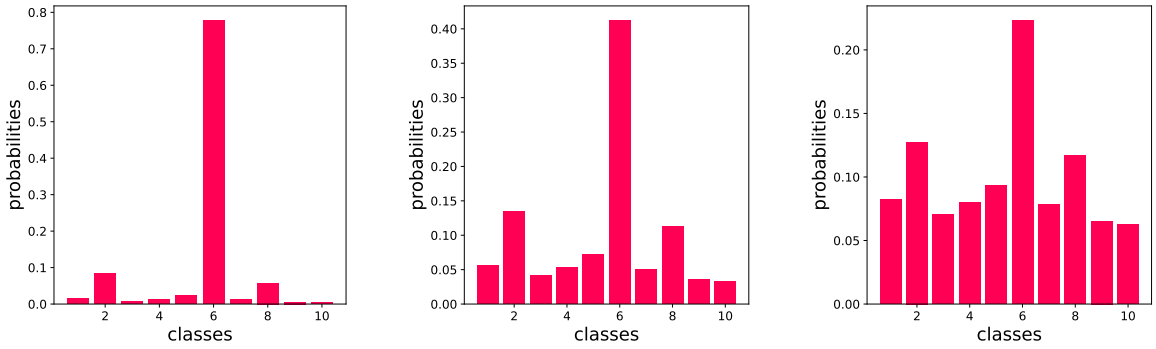
The obvious question is which layers to copy, i.e., up to which point in the source network the pre-trained weights are helpful. To address this question, Jang et al. [131] proposed a mechanism based on meta-learning to learn exactly *what* to transfer to *where*, even if the source and the target networks have different architectures. It learns meta-networks that select source-target layer pairs to transfer jointly with the training of the target network. Even if this approach benefits the model performance, the common practice usually follows the simpler pre-training

with fine-tuning.

Transfer learning can be done on all sorts of tasks provided that a source task exists. Menegola et al. [132] investigated the influence of transfer learning for melanoma screening, analyzing the usage of consecutive transfer schemes, the influence of task similarity in the results, and the benefits of transfer learning for low-resources tasks. The results confirmed a few expectations, such as performance improvement with transfer learning. However, it presented a few surprises, such as less related tasks (ImageNet to melanoma) leading to better results than related tasks (retina to melanoma) and simple transfer yielding better results than transfer pipelines. In speech processing, transfer learning is used for automatic speech recognition to adapt the acoustic model from a high-resource language to a second language, reducing the demand for data and computational resources [133]. Wang and Zheng [134] surveys other transfer learning approaches for speech and language processing.

A popular alternative to transfer learning is known as knowledge distillation, initiated with Buciluǎ et al. [135] and formalized by Hinton et al. [136]. When this approach was first introduced, the traditional method for achieving better performance on classification tasks was making an ensemble of classifiers trained with different data or different architectures and then averaging their performances. This approach works well in the laboratory environment. However, it can bring sub-optimal performance in production due to latency requirements, the large size of the networks, and the heavy memory consumption. The solution in restricted production settings is to work with a smaller model, which would often lead to lower performance. Instead of training a small model with the traditional approach, Buciluǎ et al. [135] demonstrated how to transfer the knowledge of an ensemble of large models, often referred to as the teacher model, to a single small model, often referred to as the student model, improving the generalization capability. The most straightforward approach to such a problem is to train a small model to mimic the behavior of the large one using the large model's class probabilities as "soft targets".

Nevertheless, large classification models produce very discriminative probabilities, with high confidence for their final prediction. Then, an informative statistic becomes the probability ratio between the remaining (non-predicted) classes. An illustrative example provided by Hinton et al. [136] is the case of `MNIST` digits classification. When a model predicts a "2" with 0.99 probability, it might give a probability of 10^{-3} to the class "3" and a probability of 10^{-7} to the class "7" (or the other way around). These small probabilities give an understanding regarding if that data point corresponds to a "2" that looks like a "3" or that looks like a "7". The solution given by Buciluǎ et al. [135] is to minimize the error in the logits level (before the softmax) instead of the class probabilities. The solution from Hinton et al. [136] is a generalization of



(a) $T = 0.5$ (b) $T = 1.0$ (c) $T = 2.0$

Figure 3.7: Effect of different temperatures in the softmax function.

this approach that uses high temperatures on the softmax to have smoother probabilities. The softmax function transforms logits z_i into a probability q_i as

$$q_i = \frac{\exp(\frac{z_i}{T})}{\sum_{i=1}^{N} \exp(\frac{z_i}{T})}$$

where N is the number of classes and T is the temperature – $T = 1$ is the regular softmax. A higher temperature will produce a smoother probability distribution, as seen in Figure 3.7. Hinton et al. [136] demonstrated that comparing logits is equivalent to distillation with a high temperature. In addition to the high-temperature-based training, Hinton et al. [136] includes the traditional cross-entropy loss using the true targets and predicted probabilities, showing that the mixture of the two losses improves the final performance.

The teacher-student model proposed by Hinton et al. [136] is known as response-based knowledge transfer since it tries to make the student model mimic the output probability distribution from the teacher model. Due to its simplicity, response-based knowledge transfer has been widely adopted in research. Wang et al. [137] followed a similar approach to perform semi-supervised learning with pseudo-labels for image segmentation. The intuition used in this work was that even though the model might be uncertain about the actual class of a given pixel in the image, it might be very confident regarding what *is not* the class, i.e., the network might be uncertain if the object in a picture is a dog or a cat, but it is confident that it is not a truck. Therefore, the method uses the teacher's probabilities of negative classes as pseudo-labels for a student model.

Shahrebabaki et al. [138] uses a response-based teacher-student model to perform articulatory information transfer and improve the phoneme recognition on the TIMIT dataset. Articulatory information provides essential information regarding speech; however, it is not usually available and requires specialized machinery to collect. To address the problem, Shahrebabaki

et al. [138] trains a teacher articulatory model with the "Haskins Production Rate Comparison" (HPRC) using phonemic features, which is then used to provide the articulatory features of the TIMIT data. The student model performs articulatory inversion, using the articulations from the teacher model as targets. Including the articulatory features learned by the student model serves as additional information for the phoneme recognition task and improved the performance by 6.7% on the TIMIT test set.

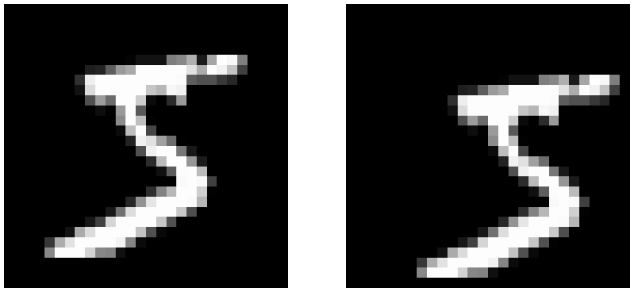
The disadvantage of response-based is that it is limited to supervised learning and does not permit intermediate-level feature learning. The alternative to overcome this problem is known as feature-based knowledge transfer. Deep neural networks are over-parameterized models, meaning that an infinite number of solutions fit the data, and the intermediate-level representations are redundant. Feature-based knowledge transfer was first introduced by Romero et al. [139], and it takes advantage of this idea to reduce the size of the fitting function. The teacher's activations in the intermediate layers are used in the supervised training of the student model in an attempt to learn similar representations with fewer parameters. A third method, relation-based knowledge transfer, explores relationships between different layers or data samples. These methods vary considerably, exploring multiple possible relationships. Yim et al. [140] uses inner products between features from different layers. Lee et al. [141] uses singular value decomposition to extract relations in the feature maps. Zhang and Peng [142] explains a multi-teacher approach for video classification.

Transfer learning and knowledge distillation are vast research fields, with several techniques and applications in most machine learning domains. For an extended guide on transfer learning, we refer the reader to Zhuang et al. [143]. For an extended guide on knowledge distillation algorithms and teacher-student architectures, we refer the reader to Gou et al. [144].

3 Deep Learning Methods for Image Processing

The multilayer perceptron was initially applied to a large set of problems, computer vision included. Consider the handwritten recognition problem, which consists of recognizing the ten digits from the `MNIST` dataset [98]. The images from `MNIST` are 28×28 pixels wide. Building an MLP for handwritten-digits recognition is as simple as flattening the image into a vector of 784 features. Then this tensor is the input of an arbitrarily large MLP that outputs the softmax probabilities over ten classes.

Since the `MNIST` dataset is standardized, this approach might achieve very reasonable performance even in the test set. However, simple data augmentation during the test will expose the MLP's flaws in computer vision problems. Consider the case of Figure 3.8a and Figure 3.8b,



(a) Sample of digit "5" from `MNIST`

(b) Translated sample of digit "5" from `MNIST`

(c) Flat features from the original digit "5"

(d) Flat features from the translated digit "5"

Figure 3.8: (a) Original `MNIST` sample, (b) `MNIST` sample after translation, (c) original `MNIST` feature vector, (d) `MNIST` feature vector after translation.

which contains the same `MNIST` digit, but with a tiny translation in the pixels. It is evident to any reasonable observer that the class is not changed; both images present the digit 5. Nevertheless, even though the images are semantically identical, the feature vector of the original input (Figure 3.8c) is entirely different from those of the new image (Figure 3.8d). The feature vector used to encode the input data does not account for the spatial relationships and the correlations between the pixels. Data augmentation during the training phase would partially address the problem. However, it would be impractical in the majority of applications.

Therefore, when we have a spatial relationship between the raw features, we need a feature extractor invariant to deformations such as translation, rotation, and others. Understanding these flaws, in 1998, Yann LeCun proposed LeNet [97], an architecture for solving the handwritten digit recognition task from the family of convolutional neural networks that we will discuss next.

3.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a family of neural networks usually applied to treat data where the spatial positioning is relevant, thus being especially interesting for image processing. The goal of CNNs is to build a feature extractor that is robust to translation, i.e., the image features are invariant to their location in the image. Robustness to other transformations, such as rotation and deformation, are desired but are not intrinsic to this family. Usually, these cases are handled by data augmentation.

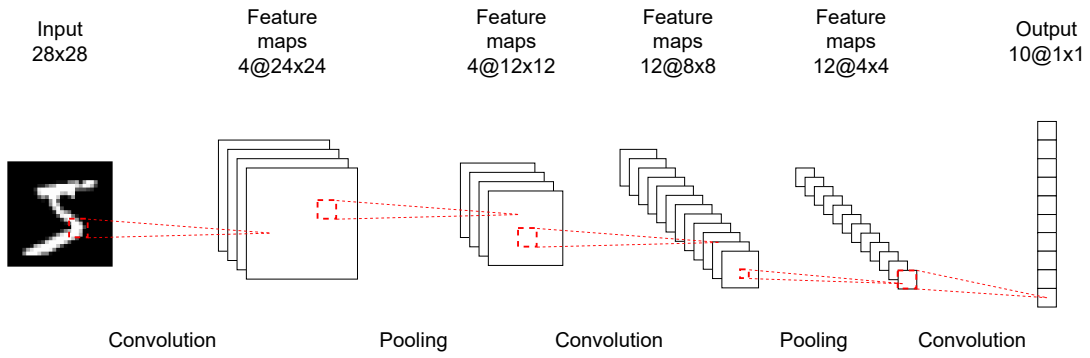


Figure 3.9: LeNet architecture for handwritten digit recognition. Reproduced from LeCun et al. [97].

The elementary operation of the CNN is convolution. Instead of performing a linear combination of the pixels, the layer weights are convoluted with the image, extracting first-order features, e.g., edges, lines, and curves. The following layers do the same but combine the first-order features building more complex ones, e.g., corners and circles. The process repeats in a way that the deeper network layers have higher-level image features, e.g., ears, eyes, wheels, and legs, composing a meaningful image representation that is used in all kinds of tasks, such as classification, segmentation, object detection, and others.

Figure 3.9 illustrates the architecture of LeNet [97], used for handwritten digit recognition. Differently from the MLP, the network takes positioning and the surrouding pixels into account. At each layer, the image resolution is reduced through an operation called *pooling* and the depth (number of channels) is increased, generating a feature space that contains finer representations. The last layer (classifier layer) is a convolutional layer in the original version. In recent architectures, the classifier is implemented as a fully-connected layer.

Even if the convolutional networks were promising, it was only by 2012 that they grew in popularity, when the AlexNet [99] conquered the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [100] by a 10.8% margin compared to the second place. AlexNet was a breakthrough when it was released. Composed of five convolutional layers and three linear layers to output probabilities for the 1 000 classes in the ImageNet dataset, it implemented many approaches that are considered standard nowadays such as non-saturating non-linearities, e.g., `ReLU`, *in lieu* of saturating non-linearities (`tanh`), training on multiple GPUs, data augmentation, dropout, and others.

The following year was marked by the development of newer and more accurate models, largely encouraged by the ILSVRC, the most relevant benchmarking so far. Many approaches have been explored to improve networks' performance, such as increasing the network's receptive field by increasing the convolutional stride. However, one of the most fruitful was increasing the

depth of the networks. Simonyan and Zisserman [145] proposed the family of networks known as VGG and investigated the impact of increasing the number of convolutional layers while keeping a small kernel size (3×3). The intuition was that even though a smaller convolutional filter would have a small receptive field, by stacking more layers, the effective receptive field would be substantially increased, with the advantage of the increased number of non-linearities and the reduced number of parameters concerning larger convolutional filters. Concurrently, Szegedy et al. [146] proposed the family of Inception networks, with its most prominent instance named GoogLeNet, referencing the LeNet architecture discussed before. The basic ideas of the Inception architecture are finding out how an optimal local sparse structure in a CNN can be approximated and covered by dense components and applying projections to reduce the dimensionality wherever the computational requirements increase substantially [146]. The VGG and the GoogLeNet achieved leading performances in the ILSVRC 2014, consolidating the literature for deeper architectures.

Szegedy et al. [147] proposed a few design principles for neural networks to improve performance, which are (1) avoiding information bottlenecks, especially in the early network layers, (2) preference for higher dimensional representations, which are easier to process by the network, (3) spatial aggregation in lower dimensional embeddings, and (4) balancing the network's width and depth. These design principles refer mainly to the Inception architecture, but they apply to most neural network families. Szegedy et al. [147] also proposes label smoothing as a regularization approach and the Inception V3, an evolution of the Inception network using convolution factorization. The final architecture is more efficient than GoogLeNet and VGG while being 42 layers deep.

In contrast, deeper architectures are more challenging to train than shallow ones. As networks become deeper, vanishing/exploding gradients arise. The problem has been addressed by different levels of normalization [148, 149, 150], enabling efficient backpropagation. However, with the increased depth, the accuracy saturates and starts degrading [151], a problem due to optimization issues rather than overfitting. To facilitate the optimization of training deeper networks, He et al. [151] proposed the ResNet, which uses a framework known as *residual learning*. ResNet is based on shortcut connections that skip one or more layers, bypassing the signal for these layers and allowing architectures eight times deeper than VGG. Formally, a neural network block learns a mapping $\mathcal{H}(x)$, where x is the network's input. The residual network would learn the mapping $\mathcal{F}(x) \coloneqq \mathcal{H}(x) - x$. Then, the original mapping is obtained using $\mathcal{F}(x) + x$. The residual learning framework achieved state-of-the-art results in image classification in the ILSVRC 2015. It demonstrated generalization ability on detection and localization

in the ILSVRC 2015 and on detection and segmentation in the COCO 2015 competition [152].

Depth substantially impacts the accuracy of deep learning models, while ResNet has shown that shortcut connections between layers improve training efficiency, allowing more layers and yielding better performance. In this direction, Huang et al. [153] proposed DenseNet, which introduces the densely connected convolutional blocks. The basic idea of the dense block is that the l^{th} layer receives the feature maps of all preceding layers as input.

$$x^{(l)} = H^{(l)}([x^{(0)}, x^{(1)}, ..., x^{(l-1)}])$$

where $x^{(l)}$ is the output of layer l^{th}, $H^{(l)}$ is a function composed by Batch Normalization [149], `ReLU` activation and a 3×3 convolution, and $[x^{(0)}, x^{(1)}, ..., x^{(l-1)}]$ refers to the concatenation of the feature-maps produced by layers $0, 1, ..., l-1$ [153]. Stacking dense blocks with pooling operations in between allows scaling deep neural networks to hundreds of layers and achieving state-of-the-art results in most image recognition benchmarks.

The presented architectures set the bases of deep learning for computer vision, bringing significant advances in various industries and particularly impacting medical image analysis [154, 155]. The state of the art set by ResNet and DenseNet was a turning point for deep learning since the challenge shifted from achieving better results in benchmark datasets to producing more efficient models, improving generalization, and requiring less annotated data during training. In this sense, Howard et al. [156] explored Neural Architecture Search (NAS) [157] in the construction of a network named MobileNetV3 that fits mobile phone CPUs. Tan and Le [158] also used NAS to design a new neural network family named EfficientNets, which are smaller and faster than traditional approaches.

The Transformer architecture [159] (which will be discussed in Section 4) was a revolution in natural language processing (NLP). However, the application of attention mechanisms to computer vision was restricted to using it in conjunction with the convolutional layers [160, 161] or replacing entire convolutional blocks [162]. Following the success of Transformers in NLP, Dosovitskiy et al. [163] proposed the Visual Transformers using an utterly different approach from the traditional CNNs. Visual Transformers divides the image into small patches (16×16 pixels), which are treated equivalently to tokens in NLP, and sequentially feed its embeddings into the Transformer. The supervised training of Visual Transformers shows that the approach performs worse when trained with mid-size datasets than the ResNet with similar size. In this configuration, the model lacks essential characteristics guaranteed by CNNs, such as locality and invariance to translation. However, when trained with much larger datasets, the visual transformers learn these inductive biases, achieving excellent results in classification benchmarks.

The models presented here were initially designed for image classification. However, they have been frequently used for many other tasks as feature extractors over the years. The basic structure of the CNNs retrieves information from the input data effectively, and the internal representation is helpful in different ways. Therefore, these architectures are the backbone for many other image-related networks, as we will see next with image segmentation.

3.2 Image Segmentation

Image recognition represented the main task in computer vision for many years. The development of deep convolutional networks drove the field to the frontier of human-level performance. The natural path would be the proposal of finer-level tasks. Object detection proposes finding the smallest envelope containing an object in the image and its label. Then, image segmentation proposes to yield a prediction for every pixel in the image, i.e., each pixel is assigned to the label of its enclosing object.

Long et al. [164] extended the success of classification networks to segmentation by reinterpreting them as fully-convolutional networks (FCNs). By combining deep, coarse, semantic, and shallow features, the FCNs resolve the dilemma between semantic and location information. While global information will encode what is in the image, the local information will resolve where it is. To convert the dense predictions into a probability map at the pixel level, the FCNs discard the linear classifiers that follow the feature extraction layers and replace them with convolutions that will restore the original image size giving a probability for each pixel. The proposal achieved leading results in the PASCAL VOC [165] 2011-2012 benchmark, which includes 20 object classes.

So far, large annotated datasets have been a strict requirement in deep image processing. Deep neural networks have millions of parameters to be adjusted, and insufficient data will inevitably incur overfitting, even with regularization, data augmentation, and dropout. Nevertheless, data availability is a limitation for many fields, including biomedical imaging, which significantly benefits from automatic image processing. In this regard, Ronneberger et al. [166] proposed the U-Net, one of the most impactful methods for biomedical image segmentation. The U-Net is an encoder-decoder architecture inspired by the FCNs literature. The network is composed of a contracting path that extracts the image features and a symmetric expanding path that reconstructs the input resolution. The features from the contracting layers are concatenated to those from their respective expanding layers as shortcut connections to improve the localization, forming a U-shaped architecture. This approach requires much fewer data points

for the training and won the ISBI cell tracking challenge 2015 [1] by a large margin. The U-Net architecture and its variants were very successful in biomedical imaging, basing much research in the future, even years after its release [167].

Chen et al. [168] introduced the DeepLab family of neural networks by combining fully connected Conditional Random Fields (CRF) [169] to the output of DCNNs, improving the overall localization of the objects and producing sharper segmentation results. The system evolved to DeepLab V2 by incorporating novel techniques: atrous convolutions and atrous spatial pyramid pooling (ASPP). Atrous convolutions use dilated convolutional kernels, allowing the control of the resolution at which features are extracted, increasing the network's receptive field without increasing computational complexity. Conversely, ASPP permits computing the convolutions in different resolutions, extracting very localized features and the overall image context. DeepLab V3 [170] further improved the system by rethinking atrous convolutions at multiple rates and batch normalization layers that proved essential for the training. In its last version, DeepLab V3+ [171] extends DeepLab V3 by adding a decoder module to refine the segmentation at the objects' boundaries. The DeepLab V3+ model was very successful in many contexts, achieving the highest performance on PASCAL VOC 2012 and Cityscapes [129] datasets, as well as proving itself useful for medical image segmentation [172].

He et al. [173] moved further by combining object detection and instance segmentation into a single network. The Mask R-CNN extends the Faster R-CNN framework [174], designed for object detection, by outputting a binary segmentation mask together with the existing object label and bounding box. The idea derives from the multi-task learning framework [175], where the same backbone is trained to perform multiple related tasks simultaneously. The joint tasks are expected to produce more general representations that would not be possible within the single-task environment. The proposed Mask R-CNN is a flexible framework that is easy to extend and adapt to different tasks, being successfully tested for human pose estimation, and showed top results in all tracks of the COCO challenges [152], instance segmentation included.

More recently, Kirillov et al. [176] introduced the Segment Anything Model (SAM), following the trend of *foundation models* [177] that is revolutionizing AI with zero-shot and few-shot generalization [91]. SAM implements *prompt engineering* to generalize to different contexts and data distributions. The model comprises an image encoder, a prompt encoder, and a mask decoder. The prompt is a set of keypoints, a bounding box, or a selection mask. A few experiments with free-text prompts, common in foundation models, were also included. SAM was released with the SA-1B dataset, containing 11M images and 1B segmentation masks,

[1] https://biomedicalimaging.org/2015/program/isbi-challenges/

configuring the most extensive image segmentation dataset to date.

Kirillov et al. [176] establishes a novel task for image segmentation called the Segment Anything (SA) task. The SA task takes inspiration from NLP to develop a prompt task for image segmentation. Given an image and a prompt, the model has to output a *valid* segmentation mask, even if the prompt is ambiguous. A valid segmentation mask means a mask corresponding to at least one object that the prompt could reference, e.g., a t-shirt or the person wearing it. The SA task substantially differs from the traditional segmentation schemes, which specialize in a set of predefined classes that can be segmented. The advantage is the simplicity of transfer to downstream segmentation tasks.

The challenges that concern image segmentation are beyond neural network architecture design. This branch of the literature is close to saturation since the performance in the most relevant benchmarks has already surpassed the human level. Still, there are obstacles to overcome. Image segmentation suffers from substantial inter-annotator disagreement, meaning that when different annotators are asked to demarcate an object in the same image, their opinions about which pixels belong to the object might diverge considerably. Ribeiro et al. [178] estimated the Cohen's Kappa score [179] for three relevant skin lesion segmentation datasets and found a median value between 0.71 and 0.75. Fortunately, Ribeiro et al. [172] showed how simple image processing approaches can minimize their impact in training deep learning models and even improve the segmentation performance concerning the ground truth annotations.

For a further review of the image segmentation state of the art, including architectures, relevant datasets, and domain adaptation, between other topics, we refer the reader to Mo et al. [180].

4 Deep Learning Methods for Sequence Learning

Sequence learning refers to techniques designed to deal with length-varying sequential data in which each sample is correlated with its neighbors. Therefore the outcome for time step t is conditioned on its context, given by the preceding $(t-1, t-2, ...)$ and, possibly, the succeeding $(t+1, t+2, ...)$ time steps, meaning that the task of a sequence model can be defined by learning the distribution

$$\mathrm{P}(x_{t+1} \mid x_t, x_{t-1}, ..., x_0)$$

in the case where the system is conditioned only on preceding time steps, or

$$\mathrm{P}(x_{t+1} \mid x_T, ..., x_{t+3}, x_{t+2}, x_t, x_{t-1}, x_{t-2}, ..., x_0)$$

in the case where the system is conditioned on preceding and succeeding time steps, where T is the sequence length.

Sequential data are often modeled as a stochastic process. A stochastic process is a collection of random variables $\mathrm{X} = \{X_t;\ t \in \{1, 2, ..., T\}\}$ defined on a probability space taking values in the state space S.

Markov processes and Markov chains are classes of stochastic processes that satisfy the limited horizon assumption, i.e., the probability distribution of the next state depends only in the current state. In general, a k^{th}-order Markov process is a stochastic process in which the probability distribution of the next state depends on the preceding k states. A stochastic process is a Markov process if it has the following properties:

1. The number of possible states is finite;
2. It is stationary, meaning that the states change over time but the probabilities governing the process remains the same;
3. The probability distribution of the following state depends only on the current state, i.e.,

$$\mathrm{P}(X_{t+1} = s_{t+1} \mid X_t = s_t, X_{t-1} = s_{t-1}, ..., X_0 = s_0) = \mathrm{P}(X_{t+1} = s_{t+1} \mid X_t = s_t)$$

for any $t \in \{1, 2, ..., T\}$ and $s \in \mathrm{S}$.

Considering the set of N states $\mathrm{S} = \{s_1, s_2, ..., S_N\}$, a Markov chain is a process that starts in a given state s_i and moves to a given state s_j with a transition probability p_{ij} [181]. Then, the transition matrix $\mathcal{P} \in \mathbb{R}^{\mathrm{S}\times\mathrm{S}}$ is defined such that

$$p_{i,j} = \mathrm{P}(X_{t+1} = s_j \mid X_t = s_i)$$

for all $i \in \mathrm{S}$ and $j \in \mathrm{S}$ [182].

The definition of Markov process permits defining the Hidden Markov Models (HMMs). The HMM is a class of generative probabilistic models that comprises two distinct stochastic processes: one defines the transitions between the states in the state space and the other defines the emission of symbols from a vocabulary V which depend only on the current state. The first is a traditional Markov chain. Therefore, the transition probabilities follow the Markovian

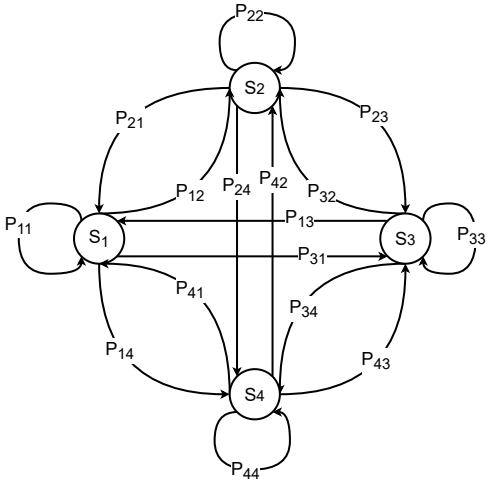


Figure 3.10: Markov chain with four states.

properties defined before. However, the states are not observable, hence, hidden. Only the sequences generated by the governing Markov process are visible. The transition and emission probabilities can be inferred from the sequences [181].

Rabiner and Juang [183] defines the HMM as a quintuple $(\mathrm{S}, \mathrm{V}, \pi, A, B)$, where

- $\mathrm{S} = \{s_1, s_2, ..., s_N\}$ is the state space with N states;
- $\mathrm{V} = \{v_1, v_2, ..., v_M\}$ if the vocabulary with M symbols;
- $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$ is the initial probability distribution on the states, i.e., π_i is the probability of the system starting at state i;
- $A = (a_{ij})_{i \in \mathrm{S}, j \in \mathrm{S}}$ is the probability of transitioning from state i to state j;
- $B = (b_{ij})_{i \in \mathrm{V}, j \in \mathrm{S}}$ is the probability of emitting the symbol j in the state i.

This model is particularly useful when there is no information regarding the states of the process and only the output sequences are available. Finally, the HMM can be used to generate a sequence of observations $O = \{o_1, o_2, ..., o_T\}$ with $o_t \in \mathrm{V}$ using Algorithm 1. HMMs have applications in several fields such as economics [184], signal processing [185, 186], speech recognition [187], and speech synthesis [188].

4.1 Recurrent Neural Networks

Unlike feed-forward networks (MLP) and CNNs, recurrent neural networks (RNNs) take variable-length inputs, meaning that they consider the present data point and its neighbors. It keeps an internal state that retains contextual information in encoding incoming information. The

Algorithm 1 Algorithm to generate a sequence of observations with a Hidden Markov Model. Reproduced from Franzese and Iuliano [181].

$t \leftarrow 1$
Sample π to obtain the initial state $S(t = 1) = s_i$
while $t \leq T$ **do**
    Sample b_{ik} in state $S(t) = s_i$ to obtain $o_t = \mathrm{V}_k$
    if $t \leq T$ **then**
        Sample a_{ij} for the state $S(t) = s_i$ to obtain the next state $S(t + 1) = s_j$
    end if
    $t \leftarrow t + 1$
end while

hidden state can store much information about the past efficiently. It can be computed in complex ways due to the non-linearities in the network design, increasing the representative power of these models.

The simplest RNN design, named vanilla RNN, uses the input at the current time step x_t and the hidden state in the previous step h_{t-1} in the computation of the hidden state at the current step h_t. For an input sequence $x = (x_1, x_2, ..., x_T)$, the RNN updates will be [189]

$$h_t = \sigma_h(W_h x_t + U_h h_{t-1} + b_h)$$
$$y_t = \sigma_y(W_y x_t + U_y h_t)$$

where y_t is the output sequence $y = (y_1, y_2, ..., y_T)$. W, U, and b are the weight matrices and bias vectors, which are learnable parameters. Alternatively, Jordan [190] proposed an update rule for the hidden state that is given by

$$h_t = \sigma_h(W_h x_t + U_h y_{t-1} + b_h)$$

The update rule for the output sequence remains the same as for Elman [189].

In some cases, the future elements of a finite-length sequence are immediately available, and it might be useful to use this information in the sequence processing. Then, the sequence is computed in the forward and backward passes. The hidden state in a bidirectional RNN is the concatenation of the hidden states in the both passes, i.e.,

$$h_t = [h_{\texttt{forward } t}, h_{\texttt{backward } t}]$$

The recurrent neural networks are trained with a gradient-based algorithm known as *back-propagation through time* (BPTT) [191, 192]. In BPTT, each time step is treated as a new layer given the previous step's hidden state, and the gradients are accumulated in the sequence

computation. The method can be inefficient, with each weight update being computationally expensive. In this regard, Sutskever [193] proposed a variant of the algorithm named *truncated backpropagation through time* (TBPTT).

TBPTT limits the dependency horizon to calculate the weights' updates. The algorithm has two hyperparameters: k_1 and k_2. k_1 refers to the number of forward passes before each weight update, and k_2 refers to the number of backward time steps to apply TBPTT. Classical BPTT is a particular case of TBPTT where k_1 and k_2 are set to the sequence length.

The crucial issue with vanilla RNNs is learning the long-term relationships due to vanishing and exploding gradients. The vanishing gradient occurs when the influence of each input step decay over time, causing the gradients to become very small and destabilizing the weights update. In contrast, the exploding gradients occur when the influence of each input step grows exponentially over time, leading to a prohibitive amount of time to learn long-term dependencies, or it simply does not work. Alternative learning algorithms were proposed to address the issue [194, 195, 196], but with little success. The most successful alternatives were the design of gated architectures.

Long Short-Term Memory

Hochreiter and Schmidhuber [197] introduced the Long Short-Term Memory (LSTM) network, and a few modifications were proposed by other researchers [198, 199]. The LSTM overcomes vanishing gradients with two main approaches: enforcing constant error flow through the internal states of the recurrent units and clipping gradients in specific network points that do not affect long-term learning.

Enforcing constant error flow through the recurrent network has two main issues. The first case is referred to as *input weight conflict*. When a recurrent cell j receives an incoming signal that is controlled by the weight w_{ij}, the same weight is responsible for storing specific inputs and ignoring others, meaning that the same parameter updates the current information and protects it from irrelevant information in the incoming message, causing a conflicting signal to w_{ij}. Similarly, the second case is referred to *output weight conflict*. When a recurrent cell j emits an outgoing signal that is controlled by the weight w_{jk}, the same weight retrieves contents from the cell j and prevents it from disturbing the upcoming cells, causing a conflicting signal to w_{jk}.

LSTM addresses the conflicting update signals by adding an *input gate* and an *output gate* that control the information flow within the cell. An additional *forget gate* controls the information retained by the cell. The forward pass of the LSTM cell is defined by the following

equations [197, 200]

$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
\tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

where σ is the sigmoid function and $\odot$ is element-wise multiplication. W, U, and b are the weight matrices and bias vectors, which are learnable parameters.

The LSTM unit keeps a memory (c_t) at each time step. The *output gate* (o_t) controls the amount of memory information that is exposed by the unit in the hidden state (h_t). The memory cell is updated considering the cell state in the previous time step (c_{t-1}) and the incoming signal ($\tilde{c}_t$). The *forget gate* (f_t) modulates the amount of past information kept in the memory cell. In contrast, the *input gate* (i_t) controls the amount of new information stored in the cell.

Gated Recurrent Unit

Cho et al. [201] introduced the gated recurrent unit (GRU) – a recurrent neural network with a similar mechanism to that of the LSTM but with only two gates, not including an engine to control the memory exposure in the hidden state, i.e., without the output gate. The GRU forward pass is given by [202]

$$
\begin{aligned}
z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\
r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\
\tilde{h}_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \\
h_t &= (1 - z_t) h_{t-1} + z_t \odot \tilde{h}_t
\end{aligned}
$$

where σ is the sigmoid function and $\odot$ is element-wise multiplication. W, U, and b are the weight matrices and bias vectors, which are learnable parameters.

Unlike the LSTM, the GRU does not keep an explicit memory cell. The hidden state (h_t) will be updated by interpolating the information stored in the previous time step (h_{t-1}) and the new incoming signal ($\tilde{h}_t$) through the *update gate* (z_t). In turn, the *reset gate* (r_t) controls the information stored from the hidden state in the previous states, allowing it to forget past

information completely. Due to having only two gates, the GRU has fewer learnable parameters, making it faster to train and requiring less data.

The vanilla RNNs always update their hidden states by combining the current input and the hidden state in previous time steps. Contrarily, gated recurrent networks (LSTM and GRU) add new content to the existing value. This additive mechanism guarantees that the features are not overwritten, facilitating the information retrieval on distant time steps. Also, it creates shortcut paths that bypass multiple time steps, allowing a more efficient gradient flow [202].

4.2 Attention Mechanisms

A particular case of sequence models is sequence-to-sequence, which maps an input sequence $x = (x_1, x_2, ..., x_{T_x})$ to an output sequence $y = (y_1, y_2, ..., y_{T_y})$ and the length of the input T_x may differ from the length of the output T_y, which is the most common case, e.g., neural machine translation, optical character recognition, speech recognition, acoustic-to-articulatory inversion.

When the input sequence has the same length as the output sequence, the problem is simplified due to the exact alignment between the input and the output. The prediction of the sequence $y = (y_1, y_2, ..., y_T)$ given the sequence $x = (x_1, x_2, ..., x_T)$ becomes [203]

$$h_t = f(x_t, h_{t-1})$$
$$y_t = g(h_t)$$

where f is a recurrent encoder network, e.g., LSTM, GRU, and g is a function usually designed as a neural network.

When the sequence lengths differ, the alignment issue needs to be considered. The traditional approach is building an encoder-decoder architecture in which both the encoder and decoder are RNNs. The encoder's job will be to project the input sequence into a context vector that contains the entire sequence context. The decoder's job will be to construct the output sequence from this latent representation. The prediction of the sequence $y = (y_1, y_2, ..., y_T)$

given the sequence $x = (x_1, x_2, ..., x_T)$ becomes

$$
\begin{aligned}
h_{\texttt{enc } 0} &= 0 \\
h_{\texttt{enc } t} &= f_{\texttt{enc}}(x_t, h_{\texttt{enc } t-1}) \\
c &= h_{\texttt{enc } T_x} \\
h_{\texttt{dec } 0} &= c \\
y_0 &= \mathrm{V}_{\texttt{START}} \\
h_{\texttt{dec } t} &= f_{\texttt{dec}}(y_{t-1}, h_{\texttt{dec } t-1}) \\
y_t &= g(h_{\texttt{dec } t})
\end{aligned}
$$

where $f_{\texttt{enc}}$ and $f_{\texttt{dec}}$ are the recurrent encoder and decoder networks, respectively, c is the context vector, $\mathrm{V}_{\texttt{START}}$ is known as start token in the vocabulary $\mathrm{V} = \{v_0, v_1, ..., v_N\}$, and g is a decoding function usually designed as a neural network. The decoding network will run an iterative process until it reaches a token representing the end of the sequence ($\mathrm{V}_{\texttt{END}}$).

In this design, the recurrent encoder network projects the entire input into a single context vector, creating an information bottleneck. Even if LSTMs and GRUs substantially advanced the modeling of long-term dependencies, they still struggle to learn very long sequences. Moreover, this design predicts the time steps sequentially; thus, as soon as the prediction for one step is wrong, the error is propagated up to the end of the inference procedure.

These drawbacks motivated the development of attention mechanisms. Attention is a natural concept for humans. When reading a text or listening to a song, we notice different parts of the sequence at a time. Likewise, when describing a scene, we will give different weights to each part of the scene for each part of the output description. Attention was first introduced in machine learning by Bahdanau et al. [204] to solve the issues related to long-term sequence learning. In attention mechanisms, instead of using only the hidden state at the end of the sequence, the context vector at each time step is a linear combination of the hidden states of all time steps in the input sequence. The coefficients of the linear combination are learned along the optimization process. Then, the context vector at time step t becomes [203]

$$
\begin{aligned}
e_{ij} &= a(h_{\texttt{dec } i-1}, h_{\texttt{enc } j}) \\
a_{ij} &= \frac{\exp e_{ij}}{\sum_{k=1}^{T_x} \exp(e_{ik})} \\
c_t &= \sum_{j=1}^{T_x} a_{ij} h_{\texttt{enc } j}
\end{aligned}
$$

where a_{ij} is the attention coefficient between the time step i in the input sequence to the time step j in the output sequence, e_{ij} represents the alignment score between the input token x_i and the output token y_j, and a is a scoring function. Initially, this scoring function was designed as the scalar product between the encoder and the decoder's hidden states. More recently, it is designed as an MLP.

Attention mechanisms have an utmost importance in sequence learning, impacting several machine learning tasks [205, 206, 159].

4.3 Self-Attention

The models described so far processes one time step at a time due to the reliance on the hidden state in previous time steps. This behavior limits the computational efficiency due to the impossibility of parallelizing the calculations. The alternative was introduced by Vaswani et al. [159], with an approach that relies only upon self-attention layers without any recurrent operation.

Contrary to the traditional attention mechanism presented before, which computes an alignment score between each step in the input to each step in the output sequence, self-attention computes alignment scores between different positions i and j in the same sequence. If on the one hand, the traditional attention mechanism computes the alignment as [203]

$$e_{ij} = a(h_{\texttt{dec}\ i}, h_{\texttt{enc}\ j})$$

where $h_{\texttt{enc}}$ and $h_{\texttt{dec}}$ are the encoder and decoder hidden states, respectively, on the other hand, self-attention computes the alignment as [203]

$$e_{ij} = a(h_i, h_j)$$

where h is an input sequence embedding.

Each input step x_i might be represented in *query*, *key*, or *value* in self-attention. The *query* representation serves the computation of the attention score in all time steps. The *key* representation is used to compute the attention score at the current time step. The *value* representation goes into the weighted sum that will give the origin to the output. The three are

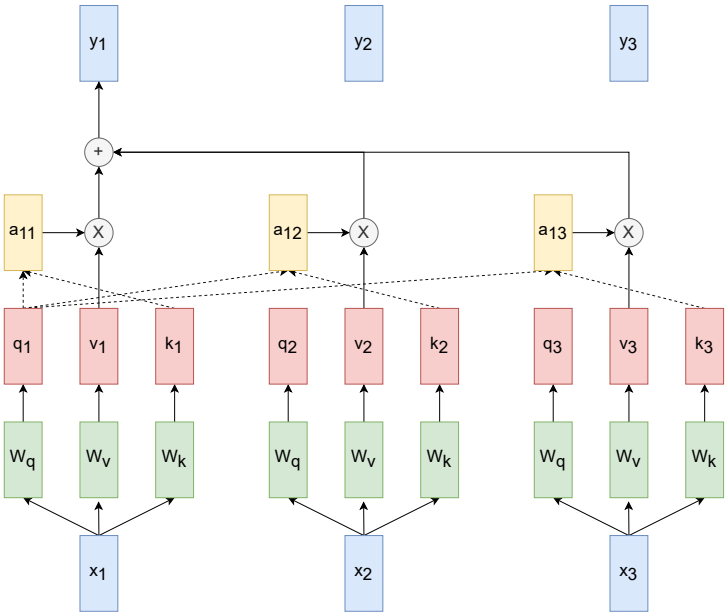


Figure 3.11: Computation of the output sequence y from the input sequence x with self-attention. Reproduced from Maucher [203].

calculated as [203]

$$q_i = W_q x_i$$
$$k_i = W_k x_i$$
$$v_i = W_v x_i$$

where W_q, W_k, and W_v are learnable parameters. Then, the attention scores and the output sequence are computed as [203]

$$a'_{ij} = \frac{q_i^\intercal k_j}{\sqrt{d}}$$
$$a_{ij} = \frac{\exp(a'_{ij})}{\sum_{k=1}^{T} \exp(a'_{ik})}$$
$$y_i = \sum_j a_{ij} v_j$$

Figure 3.11 illustrates the computation of the output sequence y given the input sequence x using self-attention. This design has two problems:

1. The query and the key for an input pair (x_i, x_j) will always be the same, but their correlation might vary in different sequences;

2. The order in the input sequence is not considered. Therefore, the output embedding will be the same regardless of the sequence order.

Vaswani et al. [159] solves the first problem using a mechanism called *Multi-Head Attention*, which introduces multiple (Q, K, V) triplets to learn different possible alignments and correlations between the input tokens. Each self-attention head has its own W_q^r, W_k^r, W_v^r matrices and produces its output sequence y^r. The final output sequence will be a combination of each head's output.

Vaswani et al. [159] also suggests a solution for the positioning problem, known as *positional encoding*. Positional encoding injects information into the input embedding about each step location in the sequence. The trivial solution would be to use the respective indices; however, the indices can grow fast in magnitude in long sequences. Normalizing the indices by the sequence length creates a new problem of sequences being normalized differently.

The original authors designed a clever approach by creating a positional encoding function that takes as input a sequence $x \in \mathbb{R}^{T \times d}$ where T is the sequence length and d is the embedding size and produces a vector with same dimensions as the input such that

$$PE(t, 2i) = \sin(\frac{t}{10\,000^{\frac{2i}{d}}})$$
$$PE(t, 2i+1) = \cos(\frac{t}{10\,000^{\frac{2i}{d}}})$$

where t is the token position in the input, and i is the embedding dimension.

4.4 Transformer

After obtaining the background in attention mechanisms, multi-head attention, and positional encoding, we can build the Transformer from Vaswani et al. [159]. The Transformer is an encoder-decoder architecture based solely on attention mechanisms, making it more parallelizable and faster to train, achieving superior quality in sequence learning tasks. Figure 3.12 presents the general architecture of a Transformer.

The encoder takes in the source sequence with positional encoding and comprises a stack of identical blocks. Each block has a multi-head attention layer followed by a fully-connected feed-forward network. It also includes residual connections around the two followed by layer normalization [207]. The decoder takes in the target sequence with positional encoding. The target is masked so the Transformer can only access the past at each time step. The decoder is also composed of a stack of identical blocks. Each block has a multi-head attention layer that learns the target sequence context. Next, a second multi-head attention layer learns the alignment between the input and the output by using the encoder's output as the key and value and the previous layer's output as the query. All layers include residual connections and are

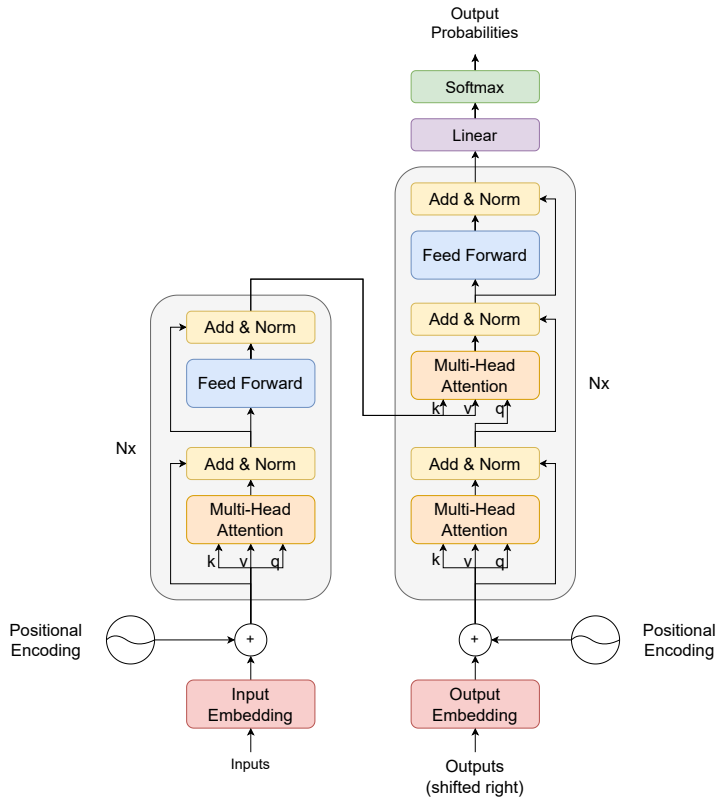


Figure 3.12: Transformer architecture. Reproduced from Vaswani et al. [159].

followed by layer normalization.

The Transformer architecture was a paradigm shift between recurrent models to attention-based architectures. Multi-head attention is used across various research fields, being the base of the most relevant language models to date, such as BERT [90] and GPT-3 [91]. We refer the reader to Phuong and Hutter [208] for a formal view of the main algorithms for Transformers.

5 Conclusion

This Chapter overviews deep learning methods that significantly impacted and continue to drive attention in general machine learning problems, computer vision, and sequence learning. Furthermore, we discussed a few relevant concepts for any deep learning problem, such as the backpropagation algorithm and knowledge transfer. Deep learning is a vast research area. We focused on the most relevant methods to this thesis; therefore, many essential aspects were left apart. Nevertheless, we took the time to go beyond the scope of the thesis when it was necessary for completeness. We referred the reader to more extensive texts for further understanding whenever possible.

Deep learning has achieved state-of-the-art results in various signal processing tasks, surpassing human-level in many of them. Nevertheless, many challenges need to be addressed.

Large models are data-consuming and computationally expensive, contributing to a significant ecological footprint [209]. Moreover, the predictions of deep neural networks are hard to interpret, raising ethical concerns and questioning if conversational AI systems are reproducing and reinforcing societal biases [210].

Despite these challenges, deep learning is a rapidly growing field revolutionizing many areas. In the following Chapters, we will explore the impact of machine learning in articulatory models of speech and show our contributions to the field.

Chapter 4

Articulatory Synthesis of Speech

> What I cannot create, I do not understand.
>
> *Richard Feynman*

1 Overview

As illustrated by Figure 4.1, articulatory synthesis is a field that has two primary research lines. The direct problem, known as *articulatory speech synthesis*, refers to synthesizing speech sounds from vocal tract articulations for a sequence of phonemes. Conversely, the inverse problem, known as *acoustic-to-articulatory inversion*, refers to estimating vocal tract articulations from speech acoustics. Articulatory speech synthesis is a challenging task, but it can potentially create more natural and expressive speech, substantially benefiting speech synthesis and conversational AI by copying the mechanisms of human speech production. An essential step for articulatory speech synthesis is the generation of realistic vocal tract articulations, a problem that we refer to as *speech articulations synthesis*.

Before deep learning, articulatory speech synthesis was regarded as the most promising approach to speech synthesis due to the ability to model extraordinary speakers, change the speaker type, alter the speech quality, and the availability of control parameters to adjust the speech and understand pronunciation mistakes [7]. The alternative so far was concatenative speech synthesis, which, even though it did not fulfill all of the desired characteristics of the epoch, was preferred by the industry due to its simplicity and reasonable quality. Articulatory speech synthesis is a complex task that requires the integration of elaborate models of the vocal tract and vocal folds, aero-acoustic simulations, and articulatory control [211], all of which should work cohesively. The level of realism of each model has a substantial impact on speech

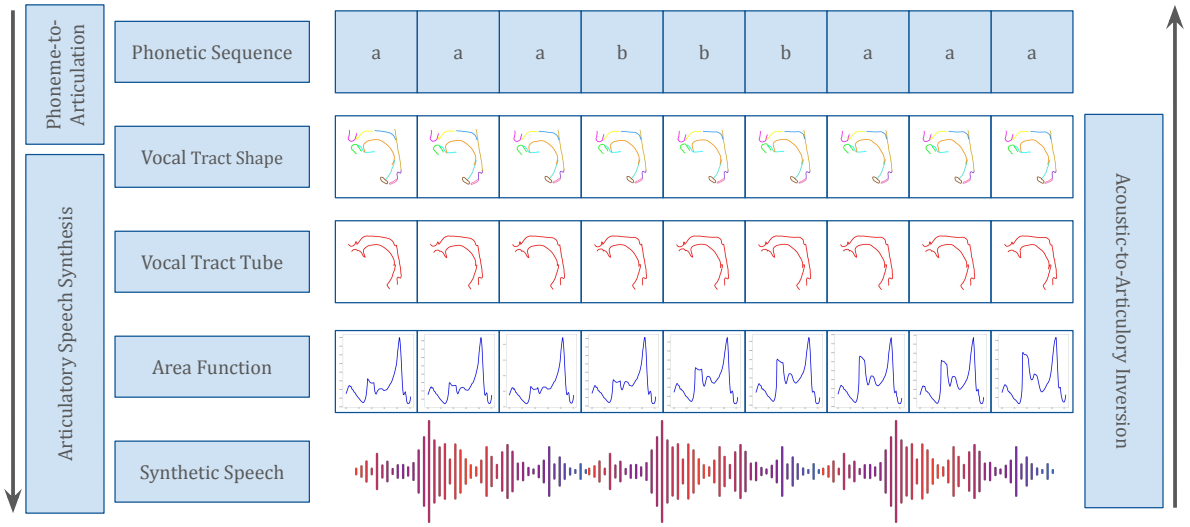


Figure 4.1: Articulatory speech synthesis and articulatory inversion illustrated.

quality and it was clear that articulatory speech synthesis was the way to achieve the desired performance. However, the growth of deep learning caused most of the traditional methods to signal processing to drop in popularity. Speech synthesis research focused on deep neural networks, which nowadays produce the most realistic speech samples in the literature [9, 212].

End-to-end speech synthesizers do not provide any information regarding the *process* of speech production. In this regard, articulatory models still hold many applications related to speech production. The capacity to understand and recreate articulatory movements from both the phonetic sequence and the acoustic signal is impactful in the study of speech, with broad applications to second language (L2) learning and speech therapy.

Describing the vocal tract movements and places of articulation is complex. Even though speaking is voluntary, it occurs automatically; usually, people are not aware of their articulations during speech. In this context, Shuster et al. [213] suggests that visual feedback is efficient to describe articulation details, especially for children. Gibbon and Wood [214] explains the usage of electropalatography (EPG) to provide feedback on tongue movements. EPG is an instrumental technique to measure tongue-palatal contact using an artificial palate containing electrodes activated by touch, providing clinically relevant information such as the place of articulation, the timing of tongue movements, and coarticulation. The paper presents the case of a nine-year-old child with abnormal speech who underwent five years of speech therapy without perceptual improvements. However, after seven once-weekly sessions with visual feedback, the subject could produce regular EPG patterns. Bacsfalvi and Bernhardt [215] shows the long-

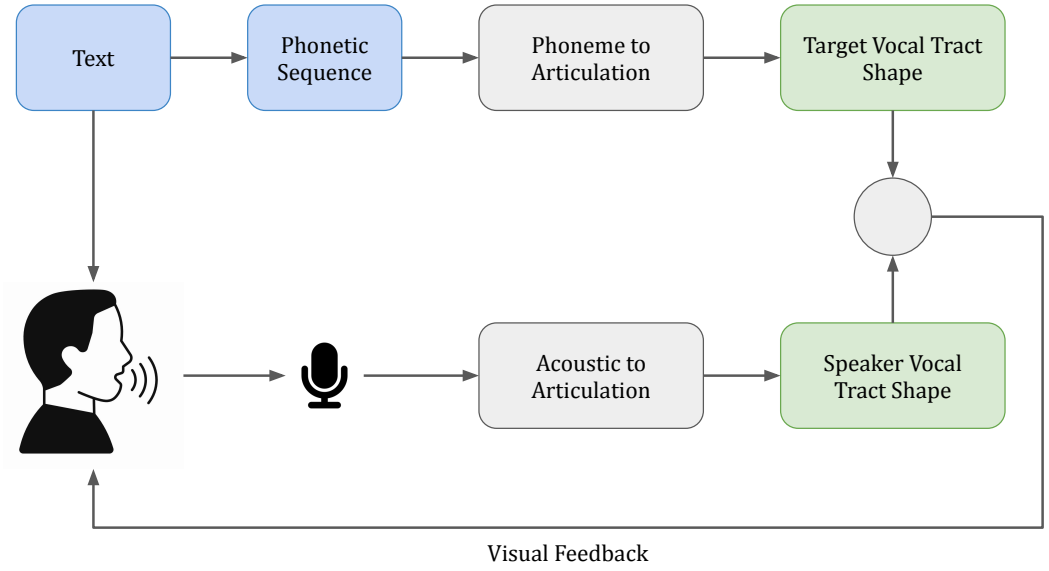


Figure 4.2: Schematic of the proposed visual feedback system. The speaker reproduces the target utterance. The phoneme-to-articulation system provides the ground truth articulation, while the acoustic-to-articulation system estimates the speakers articulations.

term outcomes of speech therapy using visual feedback. The work studies the follow-up of seven adolescents and young adults with hearing impairment after two to four years of speech therapy with ultrasound and EPG. Seven listening experts judged that five out of seven speakers maintained or continued to improve their performance post-treatment.

L2 learning can also benefit from articulatory feedback. Suemitsu et al. [216] describes an EMA-based feedback system to facilitate L2 pronunciation learning. The developed system presents real-time articulatory positions and their estimated articulatory targets. Speakers were exposed to feedback, including visual cues, acoustic cues, and visual plus acoustic cues. The results showed short-term learning effects when the visual cues were included, even without the acoustic cues. Interestingly, the acoustic cues alone did not positively affect learning.

Similarly to this study, Levitt and Katz [217] used EMA-based visual feedback on the training of eight American English speakers divided into control and treatment groups to learn a non-native consonant: Japanese post-alveolar flap. The results suggest that the speakers produced more extended flaps before the training than native speakers. However, with practice, the flap duration converged to that of Japanese speakers. Moreover, the EMA feedback benefited the learning and maintaining the novel articulation.

So far, on the one hand, the visual feedback methods used in speech therapy and L2 learning focus on the articulations performed by the speaker. However, as explained before, describing

the target movements is difficult. In this context, the direct phoneme-to-articulation model illustrates the expected vocal tract shape and the visual cues to articulate each phoneme correctly. On the other hand, together with the speaker's natural gestures, phoneme-to-articulation forms a closed feedback loop (Figure 4.2) that serves as auxiliary practice for L2 students and speech-impaired patients when they are out of the classroom or physiotherapist's office, thus impacting the learning or recovering time. In this sense, Engwall [218] proposed an articulatory tutor as a computer interface that displays the expected and the user's realizations of a set of training sounds in Swedish. The results using ultrasound data show that the virtual tutor improved the speakers' pronunciation in short-term experiments.

Nevertheless, medical equipment such as EPG and ultrasound are expensive and uncomfortable to wear. From this perspective, the acoustic-to-articulatory inversion model could more affordably replace these devices by recreating the speaker's articulatory movements from the acoustic signal, provided that it has sufficient realism.

2 Literature Review

2.1 Speech Articulation Synthesis

Speech is a dynamic and non-stationary process that requires the interaction of several articulators. It needs the rapid evolution of vocal tract configurations according to the sequence of phonemes to be articulated [219], and the phoneme articulations are very context-dependent, a phenomenon known as coarticulation [220]. Coarticulation refers to the influence of neighboring segments in articulating a given phoneme, making speech more robust to noise due to redundancy since the phonetic information is spread throughout time [221]. Coarticulation exists in two forms: backward, due to inertia, and forward, due to anticipation [222]. In this context, *articulatory control models* are those that estimate a sequence of control parameters to modulate the articulatory movements for a phonetic target, typically a series of time-labeled phonemes, which might include other markers such as prosody, emotional state, and phrasing [221]. Developing such control models faces many challenges, including the above-mentioned coarticulation effect, ensuring temporal consistency, and reaching places of articulation.

Several strategies exist to model the vocal tract dynamics during speech. Öhman [220] presented one of the first, which consists of superimposing the effect of fast constriction consonant gestures on a sequence of continuous vocalic gestures. A weight describing the degree of resistance of a consonant enables reaching the place and degree of articulation. The numerical model proposed by Öhman [223] describes the vocal tract shape $s(x,t)$ taking coarticulation

into account as

$$s(x,t) = v(x,t) + k(t)(c(x) - v(x,t))w_c(x)$$

where x is the distance from the glottis, and t is the time. $v(x,t)$ is the vowel component, $c(x)$ is the consonant component, $k(t)$ is a time-varying factor for the excursion of the consonant gesture, and $w_c(x)$ is the coarticulation function. This first model produced vocal tract shapes consistent with vocal tract measurements on X-ray images of Swedish vowel-consonant-vowel (VCV) utterances.

Alternatively to Öhman's model, Cohen and Massaro [222] interpreted coarticulation from the perspective of phonetic dominance. The authors implemented the Löfqvist's gestural production model [224] using facial articulations for visual speech synthesis with a talking head. The model considers each articulator's dominance in realizing specific phonemes and uses the negative exponential dominance function $D = \exp(-\theta\tau^c)$. The dominance falls according to the time distance τ from the center of the speech segment to the power of c; θ acting as a weight. The algorithm yields the control parameter functions for a phonetic input.

Coker [225] introduced an articulatory model that includes a physical model of the vocal structure with its spatial constraints, a representation of motion that produces intermediate shapes during state transitions, an excitation system, including subglottal pressure, vocal fold angle, and tension, and a controller that emits articulatory commands from a phonetic sequence. This articulatory model successfully described spatial and temporal characteristics of human articulation from a phonetic input reasonably matching human movements and generating intelligible speech spectrograms from simple inputs in English.

Then, Maeda [226] proposed a linear component articulatory model of the tongue and then studied the tongue's compensatory effects [31]. The temporal variations of vocal tract profiles extracted from cineradiography and labiofilm data were analyzed using factor analysis. It obtained a small set of parameters for each articulator, and the results show that these parameters can vary significantly for the vowels depending on the phonetic contexts in which they occur. Moreover, acoustic calculations show that some pairs of articulators can compensate for each other, producing similar F1-F2 patterns. These compensatory effects are most prominent for the jaw and dorsal tongue positions on unrounded vowels and jaw and lip aperture on rounded vowels. The compensatory effects of the jaw are significant since jaw opening strength considerably decreases with aging [30].

Later, Beskow [221] explored Öhman's and Cohen-Massaro's approaches for parameterizing a talking head based on phonetic input. Four models were trained to reproduce the articulatory

patterns of an actual speaker. Two are Öhman's look-ahead model and Cohen-Massaro's time-locked model, and the remaining two are based on recurrent time-delayed neural networks trained specifically for real-time applications: a symmetrical neural network and a low-latency neural network. The objective evaluation, performed over 87 test set utterances, showed that Cohen-Massaro's model yields the best matches with the ground truth regarding RMSE and correlation coefficients.

Furthermore, the symmetrical neural network demonstrated a better correlation than Öhman's model and the low-latency neural network. On a perceptual basis, 25 Swedish-speaking evaluators assessed the intelligibility. Together with the four models, the evaluation also included a rule-based approach and the audio-alone condition. The results show that although all models improved the intelligibility scores compared to the audio-alone setting, the rule-based procedure led to the highest intelligibility scores.

The difficulty for articulatory synthesis, both in the direct and inverse path, is the infinity of vocal tract shapes that can lead to the same acoustics, often referred to as the non-uniqueness problem [227]. The issue is especially relevant for neural networks since training often leads to an average solution to minimize a pre-defined cost function, missing relevant extreme positions. The answer usually requires imposing restrictions, as done by Sorokin et al. [228], which defines seven kinds of external and internal constraints; the external constraints are related to acoustics and language, and the internal are related to the vocal tract anatomy. The constraints are 1) the contractive forces of the muscles, 2) the value range for articulatory parameters, 3) mutual dependence between articulatory parameters, 4) functional dependency of the vocal tract area function on the articulatory parameters, 5) vocal tract shape, 6) acoustical deviation between recorded and synthesized speech, and 7) the complexity of planning and programming motor commands.

Some of them are unfeasible or too complex to be explored, such as the contractive forces of the muscles or the value range of articulatory parameters due to the unavailability of data that might require specialized machinery to collect, e.g., electromyography (EMG). For this reason, Potard et al. [229] took a different path, incorporating phonetic constraints derived from the knowledge of phonetics and inversion to Maeda's articulatory model [226], extensively exploring its acoustical properties. More elaborated 3D geometric articulatory models were developed [230] and then adjusted to a given speaker by exploiting a set of 3D static MRI in parallel with models derived from images.

Altogether, these models allowed a crucial scientific breakthrough in understanding compensation phenomena and its use for articulatory synthesis. However, exploiting these models faces

a significant difficulty: controlling vocal tract parameters over time. Whether a geometrical or a statistical model, the vocal tract shape is defined by a vector of parameters, which requires interpolation over time to obtain its profile at each time step. Additionally, other inputs are required for the proper synthesis of speech, such as vocal fold vibration – Elie and Laprie [231] used an improved version of the two-mass model of Ishizaka and Flanagan [232] – and temporal coordination regarding glottal opening, which plays an essential role for the production of fricatives [233] and stops. The latter requires excellent coordination between the closure at the supralaryngeal place of articulation and the glottis opening during the closure, as well as the rapid opening, which gives rise to the burst.

Articulatory phonology [234] provides a theoretical framework for representing speech using constricting events (speech gestures), which target sets of articulators, e.g., coordinative structures. The activation of these articulators corresponds to gestural scores, and their determination is the keystone of implementing task dynamics within articulatory phonology. In this regard, Nam et al. [235] exploited the XRMB database; however, the limited size explains why only the gestures' timing was obtained by warping their onset and offset to minimize the acoustic distance between natural and synthetic speech. However, adjustments to several other parameters would have been necessary. Birkholz et al. [236] proposed tenth-order linear systems to model the dynamics of articulators from a sequence of discrete consonant and vowel targets, with the expected advantage of better-fitting bell-shaped velocity profiles observed in natural movements. The parameters were optimized with an EMA database collected for a corpus of CVCVCVCV occurrences. Despite the excellent fitting of CV sequences, this work raised the issue of choosing appropriate degrees of freedom and approximating more complex phonetic contexts.

Elie et al. [237] recently presented preliminary results of an alternative to articulatory phonology proposed in Turk and Shattuck-Hufnagel [238] with a model that assumes that context-dependent targets are always reached using principles of optimal control theory (OCT). The authors define a cost function that takes intelligibility and effort into account and takes the form

$$C(\theta) = \alpha_E E(\theta) + \alpha_I (1 - I(\theta))$$

where θ are learnable parameters, E is a function of effort and I is a function of intelligibility, α_E and α_I are hyperparameters. Effort is measured as a function of the mass and velocity of each articulator. Intelligibility is measured as a function of the recognition probability of the phoneme and the duration of articulatory trajectories associated with the phoneme. The OCT-

based model permits balancing the priority between hyper-articulation (high intelligibility) and hypo-articulation (low effort).

The development of sizeable EMA databases [82] enabled the exploration of phoneme-to-articulation and articulation-to-acoustics relationships using deep learning, in direct and inverse paths, inversion being most often the focus. Richmond [239] offered the first results in the field by using a Mixture Density Network (MDN) [240], which allows estimating the probability density of the articulatory positions conditioned on acoustic features. Later, Biasutto-Lervat and Ouni [241] explored EMA to perform phoneme-to-articulatory mapping with a recurrent neural network, and Biasutto et al. [242] modeled labial coarticulation from the data collected using 44 sensors attached to an actress's face.

Csapó [243] explored RT-MRI for acoustic-to-articulatory inversion. Mid-sagittal RT-MRI images of the vocal tract were estimated using MGC-LSP spectral features as input. They showed that the LSTMs are the most suitable for the task among the evaluated models: a Fully Connected DNN (FC-DNN), a CNN, and an LSTM. Even though it is an excellent concept, the generated images contain several artifacts, and the produced shapes are not sufficiently realistic. We hypothesize that the model misused the representative capacity to learn features unrelated to speech production by targeting the complete MRI frame. Since a great portion of the image is filled with speech-invariant structures, e.g., the patient skull, and only the small region of the vocal tract correlates with the phonetic input, the model can obtain a tiny reconstruction error without producing a relevant vocal tract shape. A more efficient approach would require the model to focus only on the vocal tract, as we did in [18, 20]. In these works, the phoneme-to-articulatory mapping was done at the articulator contour level, completely ignoring any speech-invariant region of the image.

Alternatively, instead of explicitly extracting contours from the images, Gosztolya et al. [244] trained a deep autoencoder network to map articulatory features from tongue ultrasound images to acoustic spectral features. The autoencoder absorbs only the most essential image features that directly map to the targets, ignoring speech invariant features and retaining a compact set of speech parameters. The method produced relevant output, considered natural to native speakers in a listening test.

Furthermore, Attia and Espy-Wilson [245] developed a masked autoencoder approach based on bidirectional GRUs to learn articulatory patterns in the XRMB dataset, showing the capacity of these networks to reconstruct missing data up to three articulators at a time. Reconstruction of missing data is a typical self-supervision approach that permits learning the correlation of the input data, thus yielding better coarticulation models.

2.2 Articulatory Speech Synthesis

A significant challenge in articulatory research is the simulation of articulatory movements to synthesize natural and intelligible speech accounting for coarticulation. Synthesizing the acoustic signal needs solving simplified aero-acoustic equations in the synthetic vocal tract [246], which require estimating the vocal tract area function $A(x,t)$ at each time step, as done by Laprie et al. [247].

Laprie et al. [247] explored *articulatory copy synthesis* with X-ray data. Articulatory copy synthesis is a branch of articulatory speech synthesis that uses a natural utterance as a reference and varies the vocal tract control parameters until the resulting speech signal matches the expected one. The paper performed aerodynamic simulations with the model from Maeda [248] using target area functions, `F0`, and transition patterns from one vocal tract area profile to the next as input data.

The area function describes the vocal tract cross-sectional area at each point along its extension. It can be estimated using the $\alpha\beta$-model [249]

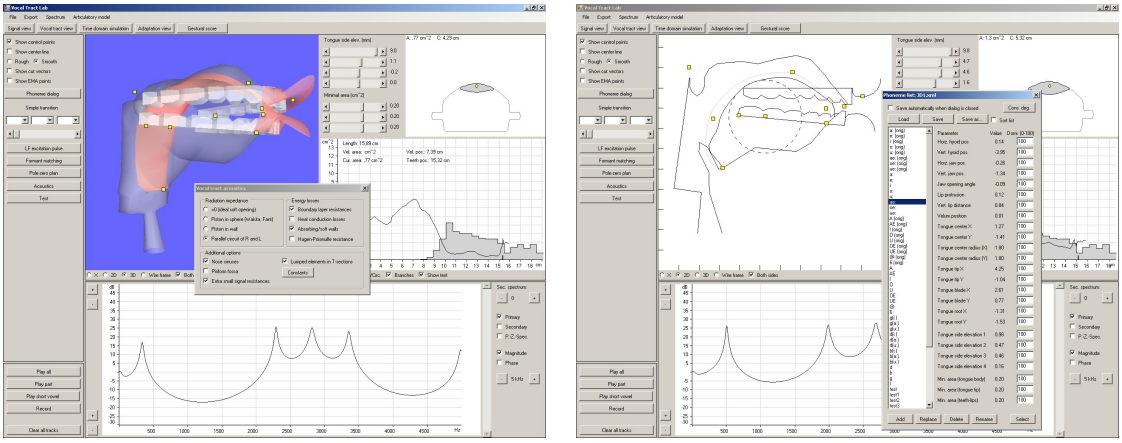
$$A(x,t) = \alpha d(x)^{\beta}$$

where x is the distance from the glottis, t is the time step, and d is the width of the vocal tract in the mid-sagittal plane. α and β are adjustable hyperparameters. In turn, the knowledge of the vocal tract center line is essential to correctly estimate the vocal tract width at the point x. Many approaches have been offered in this regard [250, 251], most being highly time-consuming. More recently, Karpinski et al. [23] proposed a neural network-based approach to accelerate this process, resulting in a solution approximately 20 times faster than the traditional approaches on average.

Alternatively, Birkholz [211] designed the articulatory speech synthesizer `VocalTractLab`, an interactive software to demonstrate the mechanisms of speech production (Figure 4.3). Birkholz [211] used the tool to synthesize CV syllables with the consonants /b, d, g, l, r, m, n/ together with eight German vowels producing isolated syllables that 50 German listeners could easily recognize in a perception test. The `VocalTractLab` synthesizer was later used by several other studies related to articulatory speech synthesis [252, 253, 254].

Using this software, Gao et al. [255] explored articulatory copy synthesis using a two-phased algorithm. The first phase uses a rule-based procedure to create an initial vocal tract position. Then, in the second phase, a genetic algorithm and a deep neural network optimize the artic-

[1] https://www.vocaltractlab.de



(a) (b)

Figure 4.3: `VocalTractLab` interface obtained from the official web page [1].

ulatory parameters to minimize the cosine similarity between the synthetic and the reference utterances.

A challenge that concerns most articulatory research is the considerable variation between speakers at the anatomical level, which usually limits approaches to a single speaker. The speaker-dependent framework has the drawback of lower data availability and the need to adapt the system for each user. In this regard, Cao et al. [256] explored speaker adaptation on the articulatory and acoustic levels to perform articulatory-to-speech synthesis using a publicly available EMA dataset [257]. The basic system design consists of an LSTM that takes as input the articulatory movements and outputs acoustic features in the form of Mel-spectrograms. Next, the Waveglow vocoder [258] converts these features into a speech waveform. Acoustic feature adaptation is performed using voice conversion [259], and articulation adaptation is done using Procrustes matching [260].

So far, articulatory research has taken mainly three directions: 1) speech articulatory synthesis, 2) articulatory-to-acoustics direct and inverse mapping, and 3) speech motor control. In contrast, self-supervision has been extensively used in speech research using large datasets to generate spoken language; however, it lacks clues regarding the speech production process. In this regard, Georges et al. [261] presents a research line in the intersection between the two, leveraging self-supervision for acoustic-to-articulatory mapping by vocal imitation. The proposed system is based on three modules: an inverse model (g), that maps acoustic features (s) to articulatory parameters (a) – $a_t = g(s_t)$; a forward model (f), that estimates acoustic features ($\hat{s}$) from the articulatory parameters – $\hat{s}_t = f(a_t)$; and a pre-trained articulatory synthesizer (ϕ) that reproduces the acoustic features ($\tilde{s}$) from the articulatory parameters – $\tilde{s}_t = \phi(a_t)$. The models learn the forward and the inverse articulatory-to-acoustic mapping

jointly by minimizing the discrepancy between the estimated signal ($\hat{s}$) and the input signal (s) together with the disparity between the estimated signal ($\hat{s}$) and the reproduced signal ($\tilde{s}$) using the accommodation algorithm proposed by Laurent et al. [262]. The results show exciting directions in using self-supervision for articulatory models, allowing the adaptation of sensory-motor architectures to actual speech.

3 Evaluation of Articulation Synthesis

The evaluation of vocal tract shape models is mainly challenging due to the nonexistence of a proper target. Even if articulatory data is available, it is impossible to say that it is the ground truth. That is because the term ground truth in machine learning usually refers to an absolute truth – a single target to pursue. Due to the non-uniqueness of speech articulations and the difficulty to define a ground truth in many medical tasks, it is only possible to say that the speech instance in hand is one of the possible solutions. The variability in the speech articulations of a single speaker uttering the same phoneme is called intra-speaker variability.

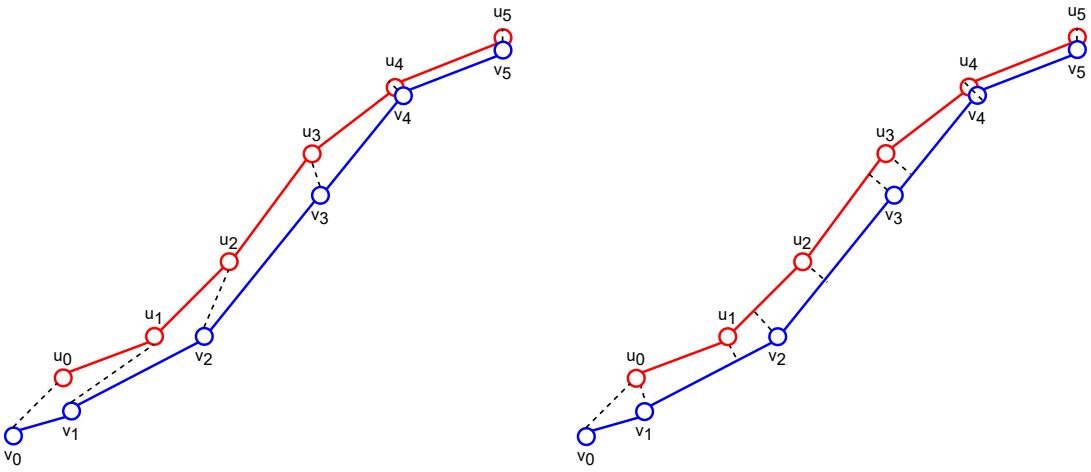
A second challenge for evaluating articulatory models are the anatomical differences in the vocal tract and the various speaking strategies between multiple speakers, called inter-speaker variability. Serrurier et al. [263] characterized speaker-independent articulatory strategies and inter-speaker variability for eleven French speakers uttering 62 vowels and consonants. The model was capable of explaining 66% to 69% of the variance. However, the research suggests that most of the variability is due to anatomical differences instead of speaking strategies. This problem usually limits the usage of multi-speaker datasets in articulatory synthesis and articulatory-to-acoustic inversion models, requiring speaker normalization [264, 265].

The typical evaluation strategies are based on point-wise error metrics and correlation coefficients. These methods usually fit EMA data well due to the small number of key points involved. The distance between the target curve $u \in \mathbb{R}^{N\times 2}$ and the predicted curve $v \in \mathbb{R}^{N\times 2}$ is visually represented by Figure 4.4a and is given by

$$\texttt{Euclid}(u, v) = \frac{1}{N}\sum_{i=1}^{N} d(u_i, v_i) \tag{3.1}$$

where $d(u_i, v_j)$ is the Euclidean distance between points u_i and v_j.

The disadvantage of this metric is that it requires perfect alignment between the target and the predictions, which is the case for EMA since the number and the anatomical location of the sensors is fixed, but does not suit geometrical curves well. Therefore, an alternative is the



(a) Euclidean distance

(b) Orthogonal distance

Figure 4.4: Representation of the euclidean and orthogonal distances between curves u and v.

orthogonal distance, which is visually represented by Figure 4.4b and given by

$$\texttt{Orthogonal}(u, v) = \frac{1}{2N} \sum_{i=1}^{N} (d(u_i, v) + d(v_i, u)) \tag{3.2}$$

where $d(u_i, v)$ is the orthogonal distance between the point u_i and the curve v; likewise, $d(v_i, u)$ is the orthogonal distance between the point v_i and the curve u. The difficulty is related to the extreme points, where the orthogonal projection is undefined. In practice, the closest point is considered, which is the case of v_0 in Figure 4.4b.

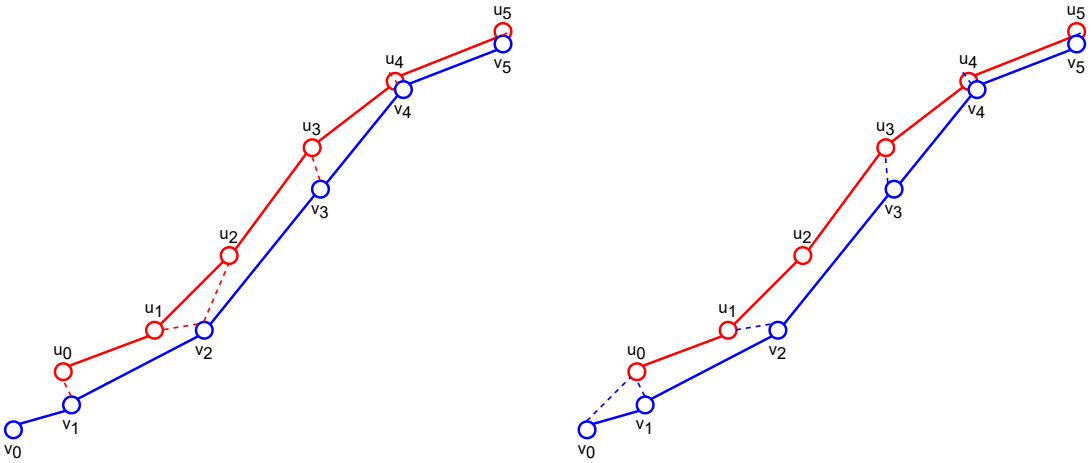
Labrunie et al. [266] proposed to use the point-to-closest-point distance (P2CP), visually represented by Figure 4.5, which is less conservative than the Euclidean distance, to evaluate the tracking of vocal tract articulators in MRI. The mean point-to-closest-point distance ($\texttt{P2CP}_{\texttt{mean}}$) between the target ($u \in \mathbb{R}^{N \times 2}$) and the predicted ($v \in \mathbb{R}^{N \times 2}$) curves is given by

$$\texttt{P2CP}(i, u, v) = \frac{1}{2}\left(\min_{j \in \{1,2,\ldots,N\}} d(v_i, u_j) + \min_{j \in \{1,2,\ldots,N\}} d(u_i, v_j)\right)$$

$$\texttt{P2CP}_{\texttt{mean}}(u, v) = \frac{1}{N} \sum_{i=1}^{N} (\texttt{P2CP}(i, u, v))$$

where $d(u_i, v_j)$ is the Euclidean distance between points u_i and v_j. Alternatively, the root mean square (RMS) is more sensitive to outliers. In contrast, the mean value is more popular in the literature [266]. The $\texttt{P2CP}_{\texttt{RMS}}$ is given by

$$\texttt{P2CP}_{\texttt{RMS}}(u, v) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\texttt{P2CP}(i, u, v))^2}$$



(a) Curve u relative to curve v (b) Curve v relative to curve u

Figure 4.5: Representation of the `P2CP` distance between curves u and v.

Point-wise error metrics provides a reasonable measure of the deviations between two curves, being easy to interpret and compute. They are the gold standard for tasks with a clear objective, such as articulatory tracking in medical images, which is the case of Labrunie et al. [266] and the experiments described in Chapter 5. However, they are not optimal for articulatory synthesis tasks due to the non-uniqueness problem, and it gives equal weight to all the points regardless of constrictions, which is unrealistic since constricted regions have a decisive impact on acoustics. In this sense, other metrics should be taken into account. It is essential to evaluate if the predictions follow a similar dynamic to the targets. These dynamics are usually computed as correlation coefficients in the x- and y-axis. The difficulty in analyzing correlation coefficients is understanding when they are relevant, e.g., the pharynx's x-axis correlations are mostly useless since they have mainly vertical movements. For this reason, correlation coefficients are most commonly used with EMA data due to the correspondence of sensors and anatomical points and their known expected dynamics.

The presented metrics focus on individualized articulators. Nevertheless, it is rather the interactions between articulators other than their positions alone that allow speech production. Therefore, an alternative for studying vocal tract dynamics during speech is through their associated tract variables. Tract variables (TVs) are measurements made at specific points of the vocal tract, representing the constriction between articulators. Figure 4.6 presents a visual representation of the TVs, and Table 4.1 presents their names and the associated constrictors. Speech gestures are intended to reach articulatory goals and are defined in terms of TVs [234].

The TV trajectories must reflect critical articulators. Critical articulators are those whose positions are imposed to achieve the target place of articulation. They resist context and have coarticulatory effects on neighboring phones [268]. For example, complete contact between the

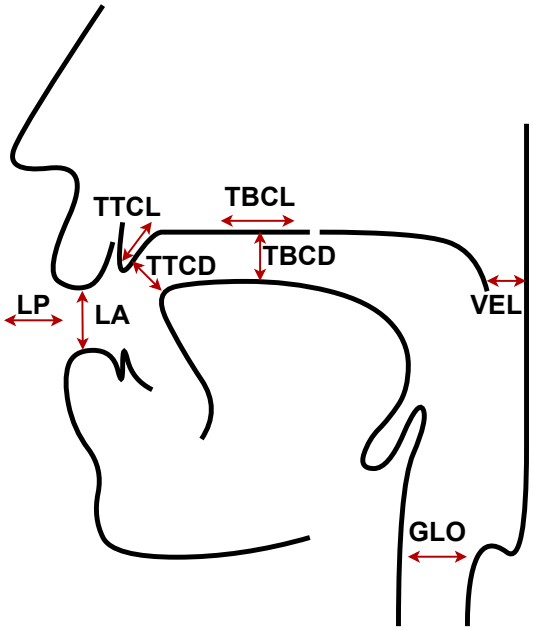


Figure 4.6: Visual representation of the tract variables. Reproduced from Saltzman and Munhall [267].

Table 4.1: Tract variables and their associated constrictors. Source: Reproduced from Browman and Goldstein [234]

	Tract Variable	Constrictors
LP	Lip protusion	Upper & lower lips, jaw
LA	Lip aperture	Upper & lower lips, jaw
TTCL	Tongue tip constrict location	Tongue tip, tongue body, jaw
TTCD	Tongue tip constrict degree	Tongue tip, tongue body, jaw
TBCL	Tongue body constrict location	Tongue body, jaw
TBCD	Tongue body constrict degree	Tongue body, jaw
VEL	Velum aperture	Velum
GLO	Glottal aperture	Glottis

lips is mandatory for labial phonemes, minimizing LA. Similarly, the tongue tip must touch the alveolar region to produce dental phonemes, minimizing TTCD. For palatals, the tongue dorsum must approach the palate, minimizing TBCD. For nasal phonemes, the velum lowers to enable airflow through the nasal cavity, increasing VEL. As an example, Kim et al. [269] studied the critical and non-critical articulations in emotional speech production, detailing the effect of emotion in speech articulations on consonant-vowel-consonant utterances.

Directly measuring the TVs is an efficient method to evaluate if the articulatory synthesizer reaches its articulatory targets. Nevertheless, it can be tricky since separating the cases where that TV is relevant is necessary. Moreover, there are cases where a constriction without contact is required. It is the case of fricatives, e.g., the lips for /f/ and the tongue and alveolar region for /s/, when a small channel without a complete closure is necessary to produce the fricative noise. The same occurs for some vowels, e.g., /i, y/, which requires a small channel in the front of the oral cavity. An alternative approach is measuring the correlation coefficients between the target and the predicted TV trajectories. By doing so, we can directly quantify if the models produce similar dynamics as those of the target shapes.

Tract variable evaluation suits consonants well. However, it does not fit most vowels, characterized by the resonator's shape and not by constrictions. A more suited metric for vowels would be measuring formant frequencies [270] by solving simplified aero-acoustic equations in the synthetic vocal tract [271], which is computationally expensive. Alternatively, speech synthesis metrics could be used, such as the mean opinion score [272]; however, we incur the same computational cost of measuring formants. The best synthesizer would be the one that outputs the most intelligible speech. External evaluators could grade the synthetic vocal tract shapes and measure the realization of the relevant articulatory features, requiring highly specialized evaluators familiar with the acoustics and dynamics of the vocal tract. However, these metrics are very subjective.

Recent research, however, has put significant attention into articulatory feature classification and its use in understanding the relationship between articulations and acoustics and how neural networks map the two. Elie et al. [237] used phoneme recognition probability as a measure of intelligibility in their cost function. Saha et al. [219] trained a Long-term Recurrent Convolutional Network to classify 51 VCV contexts from RT-MRI films from 17 speakers, obtaining an accuracy of 42%. Van Leeuwen et al. [273] trained a CNN to classify sustained phonemes (vowels and fricatives) from static mid-sagittal MRI and obtained an accuracy of 57%. Curiously, the model learned representations compatible with the vowel chart, showing that although the accuracy is limited, the model is consistent with the phonetics literature. On the problem of

evaluating synthesized vocal tract shapes, Engwall [274] used an articulatory classifier as an evaluation metric for acoustics-to-articulatory inversion of VCV words in Swedish sentences using linear estimation and neural networks. This last work shows that the articulatory classifier provides a more intelligible metric than RMS error and correlation coefficients. In Section 5 from Chapter 6, we took a similar direction by using phoneme recognition as a measure of phonetic information in the mid-sagittal contours.

4 Conclusion

This Chapter contextualized articulatory synthesis of speech, a multi-faced research field. We defined articulatory speech synthesis as synthesizing speech sounds from vocal tract articulators and its counterpart, acoustic-to-articulatory inversion, as the task of estimating the speech articulations from an acoustic signal. We also defined our main research topic, speech articulation synthesis from a phonetic sequence (phoneme-to-articulation), as the task of estimating the vocal tract articulations for each phoneme to be articulated. In this context, phoneme-to-articulation is a crucial step for articulatory speech synthesis since it maps the input phonetic sequence to its respective articulations, which can later be used to synthesize speech.

Even though end-to-end models dominate the current speech synthesis research, these approaches need more links to speech production theory, being insufficient for fields that require a precise understanding of vocal acoustics and the bio-mechanical processes. We reviewed areas in which the articulatory speech process can tremendously impact, such as L2 learning and speech therapy. Then, we discussed many traditional and recent articulatory models with a direct impact in understanding coarticulation, compensatory mechanisms, speech strategies, and others. Beyond the importance of these models in speech production theory, recent research has shown how articulatory models can improve speech synthesis and automatic speech recognition, a potential asset for many industry applications.

Lastly, we discussed the evaluation of speech articulatory models. These models are hard to measure and compare due to the large inter- and intra-speaker variability, speaker normalization, non-uniqueness, and other challenges. We presented traditional point-wise metrics helpful for vocal tract segmentation and EMA, tract variables correlation coefficients, which are useful for evaluating the dynamics and interactions between the articulators. More recently, articulatory feature classification has concentrated significant attention by enabling a mapping between articulations and acoustics. The advantage of articulatory feature recognition is the possibility of assessing speech intelligibility in a fast and computationally cheap manner. Furthermore, with the availability of multi-speaker articulatory data, the classifier can learn a

speaker-independent recognition function, which is a more efficient metric than RMS error and correlation coefficients.

Chapter 5

Automatic Segmentation of Vocal Tract Articulators in RT-MRI

> What we observe as material bodies and forces are nothing but shapes and variations in the structure of space.
>
> *Erwin Schrodinger*

1 Overview

As discussed in Chapter 2, the characterization of the complete vocal tract geometry is essential for many aspects of speech research and RT-MRI is the preferred data modality in the current literature. Still, the images alone are insufficient for many applications. The reason is that the exact geometry of the vocal tract air column, determined by contours of the articulators from the glottis to the lips, is often needed, which is the case of this thesis.

Articulatory synthesis of speech requires characterizing the vocal tract geometry in the level of speech articulators to study their individual contributions to speech. The segmentation of vocal tract articulators serves as a pre-processing step on the articulation synthesis pipeline to obtain the training data for neural networks. In this Chapter, we describe our proposal for the problem. Our method consists of two parts. The first is the segmentation of individual articulator borders using a convolutional neural network. The second uses specific rule-based algorithms to post-process the segmentation masks (network's outputs) to define the contours of the articulators.

2 Literature Review

Raeesy et al. [275] proposed a method of automatic landmark tagging in which a recursive boundary subdivision algorithm [276] extracts a set of landmarks corresponding to the vocal tract contours. Then, the oriented active shape model [277] recognizes and delineates the vocal tract shapes. However, the small dataset (25 images from five speakers) limits the significance of this work. Alternatively, Silva and Teixeira [278] proposed an unsupervised method for articulator segmentation based on a modified version of the active appearance model [279]. The technique takes advantage of the low inter-frame differences and is based on 26 vocal tract landmarks manually marked per frame in a training database of 51 images. Moving towards machine learning algorithms, Labrunie et al. [266] explored a large RT-MRI corpus in French for training three supervised segmentation methods: Multiple Linear Regression (using pixel intensities), a modified version of the Active Shape Model (mASM), and Shape Particle Filtering (using more elaborate image features). In a leave-one-out cross-validation scheme, the three methods were compared on several articulators using the point-to-closest-point distance (`P2CP`). The results showed that the mASM outperforms the other two for all articulators.

Meanwhile, deep neural networks have become the standard for computer vision and medical image processing [280]. CA et al. [281] segmented air-tissue boundaries (ATBs) using a Fully Convolutional Network (FCN) [164] followed by a canny edge detection algorithm to output smooth and realistic ATBs. Following a strategy similar to Fasel and Berry [282], Jaumard-Hakoun et al. [283] trained a deep neural network based on the stacking of Restricted Boltzmann Machines [284] with the contours extracted by an automatic algorithm that uses block-matching to enforce the frame-to-frame similarity. Eslami et al. [285] explored the segmentation of the jaw, given by the lower incisor profile, tongue, and vocal tract air cavity on static mid-sagittal MRI for ten subjects sustaining 62 articulations. The method uses a modified version of the pix2pix algorithm [286], taking advantage of the conditional generative adversarial networks.

While most of the approaches presented so far focus on a single-frame segmentation, Asadiabadi and Erzin [287] proposed a sequence-to-sequence Deep Temporal Regression Network to estimate the coordinates of the vocal tract curve and the points separating the articulators, providing individualized contours for each articulator, which is essential to study their contributions during speech. As Hebbar et al. [288] show, using temporal information instead of making single-frame predictions improves the segmentation when the articulators are in contact. Finally, Isaieva et al. [289] offered an alternative in which only the pixels in the tongue's edges are segmented. The method uses a U-Net [166] for image segmentation, followed by a graph-based algorithm (discussed later in the Chapter) to convert the soft probabilities in the

network's output into the tongue contour.

Our state of the art review revealed several gaps in the literature. Few papers provide individual contours of all non-rigid articulators. Additionally, only some studies demonstrate generalization across multiple speakers. Finally, none provide an open repository for public usage and evaluation of the proposed methods. Our work aims to fill these gaps by presenting a robust speaker-independent approach to segment the vocal tract articulator contours in RT-MRI movies. We also investigate a speaker adaptation approach to enhance the performance of a target subject. This work proposes a single deep convolutional neural network (DCNN) that can automatically trace the boundaries of nine vocal tract articulators in RT-MRI frames. The DCNN takes RT-MRI frames as input and outputs a probability map over the pixels belonging to the articulator's boundaries. Additional post-processing is then used to obtain a curve giving the exact shape of each articulator. This work is intended for articulatory speech research, including but not limited to articulation and articulatory speech synthesis. The main contributions of this work are:

- The coverage of the main non-rigid articulators necessary for speech production;
- The assessment of inter-speaker generalization through leave-one-out cross-validation (LOOCV) protocol;
- The processing of vast RT-MRI corpora with a low error in comparison to human annotations;
- The public availability of the segmentation system[1], allowing it to be tested and audited by the scientific community.

It is important to highlight that a substantial part of this Chapter is contained in Ribeiro et al. [19]. We refer the reader to the original work for additional information and supplementary material.

3 Materials

3.1 Datasets

The corpus for this research comprises two real-time MRI datasets of French speakers, `ArtSpeech Database 1` (`ASD1`) and `ArtSpeech Database 2` (`ASD2`). The `ASD1` is a part of the database described in Isaieva et al. [87]. While the published database contains ten subjects, only seven

[1] https://gitlab.inria.fr/multispeech/vt/vt_tracker

Table 5.1: Parameters of the MRI acquisition.

Parameter	Value
TR	2.22 ms
TE	1.47 ms
FOV	22.0 cm × 22.0 cm
Pixel Spacing	1.62 mm/pixel
Flip Angle	5 degrees
Slice Thickness	8 mm
Num. of Radial Encoding Lines per Frame	9
Pixel Bandwidth	1 670 Hz/pixel
Image Resolution	136 × 136 pixels

(denoted in this text as S1-S7) were used in this study because the larynx was not visible in the images of the other three. The `ASD1` contains a total of 365 400 frames. The same protocol was used to acquire a larger speech corpus with 320 000 frames from a single subject, forming a new dataset denoted as `ASD2`. The subject participating in `ASD2` acquisition was the last subject S7 from the dataset `ASD1`. To distinguish these two sets, the data of S7 from `ASD1` is denoted as S7.1, and its data from `ASD2` is marked as S7.2.

Both datasets were collected using state-of-the-art protocols and recommendations in the laboratory IADI at the Centre Hospitalier Régional Universitaire de Nancy, France. All participants provided written informed consent, and the data were recorded under the approved ethical protocol "METHODO" (ClinicalTrials.gov Identifier: NCT02887053). The study was approved by the institutional ethics review board (CPP EST-III, 08.10.01).

The images were acquired with a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany). The radial RF-spoiled FLASH sequence [65] was used, with the parameters listed in Table 5.1. The films were recorded at a frame rate of 50 fps and reconstructed with the algorithm presented in Uecker et al. [65]. During the acquisitions, the speakers were asked to read out loud sentences that were presented to them. Each acquisition took about one and a half minutes.

Due to the difficulty of annotating all of the images in the datasets, samples with a good representation of variability were selected. Initially, this selection was conducted independently for datasets `ASD1` and `ASD2`, by different annotators, using slightly different methodologies. Later the collected and annotated data were merged to increase the database. In Isaieva et al. [289], it was shown that several hundreds of hand-annotated images were sufficient for a good tongue segmentation. Therefore, sample sizes of both datasets was selected to be of this order. To ensure the best variability coverage, the images for manual annotation were selected with the k-means algorithm. For `ASD1` a $k = 100$ was selected. The k-means algorithms was applied independently for each subject, and only the closest to the cluster centers images were kept,

Table 5.2: Number of annotated samples per subject per dataset split.

Dataset	Subject ID	ID on Isaieva et al. [87]	Gender	Train Images	Validation Images	Test Images
ASD1	S1	P1	Male	71	9	20
	S2	P3	Male	71	9	20
	S3	P5	Male	71	9	20
	S4	P6	Male	71	9	20
	S5	P7	Female	71	9	20
	S6	P9	Female	71	9	20
	S7.1	P10	Female	50	0	50
ASD2	S7.2		Female	310	54	63

resulting in 700 images in total. For ASD2 the algorithm was applied with $k = 10$. The clusters were evenly sampled, resulting in 427 images.

The two datasets were split into train, validation, and test as described in Table 5.2. The data was divided at the complete sequences level, so all samples from the same acquisition were placed at the same split. The reason is that adjacent images are very similar, and putting them in separate sets would introduce bias into the train-test scheme. The validation set for S7.1 is empty because it is the same subject as S7.2, which already has a sizeable validation set.

3.2 Annotation Procedure

We performed semi-automatic annotations of the articulator boundaries. Our previous segmentation system [20] was used to track the upper vocal tract cavity, including the two lips, tongue, soft palate, and pharynx. For the laringeal articulators, which were not included in the previous study, we trained a Mask R-CNN network with the 427 samples from ASD2 described in Table 5.2 to produce a first guess of the contours for each articulator. The models were then used to automatically annotate the unlabeled images in the dataset. The automatic annotations were then carefully reviewed and manually corrected. This semi-automatic procedure allowed us to complete the annotation protocol with limited resources. The image annotations were made as follows:

- **Arytenoid Cartilage:** Through their vocal process, these cartilages are the siege of the vocal cord attachment to the posterior part of the larynx (represented by the cricoid cartilage). The complete extension of the arytenoid cartilages were annotated, covering a vertical range of about two vertebrae (at the level of the $5^{\text{th}}/6^{\text{th}}$ cervical vertebrae). The annotation started at point A in Figure 5.1a and continued to point B, passing through point C.

- **Epiglottis:** The epiglottis is a thin and elongated cartilaginous structure describing the upper-anterior part of the larynx. Given the reduced thickness of this cartilage, we chose to annotate the epiglottis center line, starting from the anterior part of the larynx to the epiglottis posterior extremity. Laprie et al. [290] provided an algorithm for reconstructing the epiglottis from the center line. The annotation started at point `D` in Figure 5.1a and continued to point `E`.

- **Lower Lip:** This part begins from the anterior part of the mandible (at the lower bottom of the gingiva vestibule) to the external lip hem. The contour of the lower lip was annotated from point `F` to point `G` in Figure 5.1a.

- **Posterior limit of the pharynx:** This area was annotated from the posterior part of the nasal cavity to the cricoid cartilage (behind the arytenoid cartilage). The annotation started at point `H` in Figure 5.1a and continued to point `A`.

- **Soft Palate:** The soft palate appears as an elongated structure in the mid-sagittal plane, similar to the epiglottis. For the same reason, we only annotated the center line of this area, from the posterior limit of the hard palate (motionless) to the posterior extremity of the moving part. Like the epiglottis, the algorithm from Laprie et al. [290] can reconstruct the soft palate from the center line. The annotation started at point `I` in Figure 5.1a and continued to point `J`.

- **Tongue**: The complete extension of the tongue (apex, dorsal part, and root) was annotated, starting at the frenulum on the mouth floor and ending the tongue at the root below the hyoid bone. The annotation started at point `K` in Figure 5.1a and continued to point `L`. The sublingual cavity was marked when visible. Since the MRI is 8 mm thick and the frame rate is low compared to the speed of tongue movements, the uncertainty related to the partial volume effect [75] and blurring is specially marked in the tongue tip and dorsum. In both cases, we decided to annotate the most visible contour.

- **Thyroid Cartilage:** This part was annotated as a closed contour, starting at the anterior limit of the epiglottis (at the level of the $3^{\text{rd}}/4^{\text{th}}$ cervical vertebrae) and ending below the vocal folds (approximately at the level of the 7^{th} cervical vertebrae). This annotation is drawn as the oval shape passing through points `M`, `D`, and `N` in Figure 5.1a. The position of the thyroid cartilage is more important than its precise shape for confirming the position of the glottis.

- **Upper Lip:** This upper part was drawn from the anterior nasal spine (at the upper

bottom of the gingiva vestibule) to the external upper lip hem ("cupid's bow"). The complete contour of the upper lip was annotated from point `O` to point `P` in Figure 5.1a.

- **Vocal Folds:** The vocal folds are not entirely visible in the MRI frame. Only the negative of the glottis is observable between the thyroid cartilage and the arytenoid cartilage. The vocal folds are marked as an oval shape passing through `D` and `C` in Figure 5.1a.

Figure 5.1 shows an MRI frame with the landmarks used to reproduce the annotation procedure and three MRI samples with superimposed annotations for each articulator. A dental surgeon with seven years of experience validated the annotation procedure.

4 Methods

The strategy used to track the shapes of the articulators is similar to that used by Isaieva et al. [289] for the tongue. It comprises two phases:

1. A deep convolutional neural network (DCNN) is trained to estimate the probability that a pixel belongs to the articulator's contour;
2. A post-processing algorithm is applied to the network's outputs (segmentation mask) to construct the curve describing the articulator's shape. The nature of the algorithm and its hyperparameters depend on the articulator.

Section 4.1 describes the learning strategy used in the first phase, while Section 4.2 describes the algorithms applied to each articulator. The final contours are regularized using b-splines to smooth the output and match the target and prediction lengths[2].

4.1 Articulator Boundary Segmentation

Unlike Isaieva et al. [289], who used the U-Net [166], we chose to work with the Mask R-CNN [173]. Mask R-CNN is a simple and flexible framework for object instance segmentation. It is lightweight, easy to train, and can be applied to different tasks. These characteristics, as well as the availability of a pre-trained implementation[3] on standard deep learning libraries, make Mask R-CNN one of the preferred methods for medical image segmentation.

The Mask R-CNN architecture is advantageous for our problem because it performs three tasks simultaneously: object detection, classification, and segmentation. This approach allows

[2]The b-spline regularization function can be found at https://gitlab.inria.fr/multispeech/vt/vt_tools/-/blob/main/vt_tools/bs_regularization.py

[3]https://pytorch.org/vision/main/models/generated/torchvision.models.detection.maskrcnn_resnet50_fpn.html

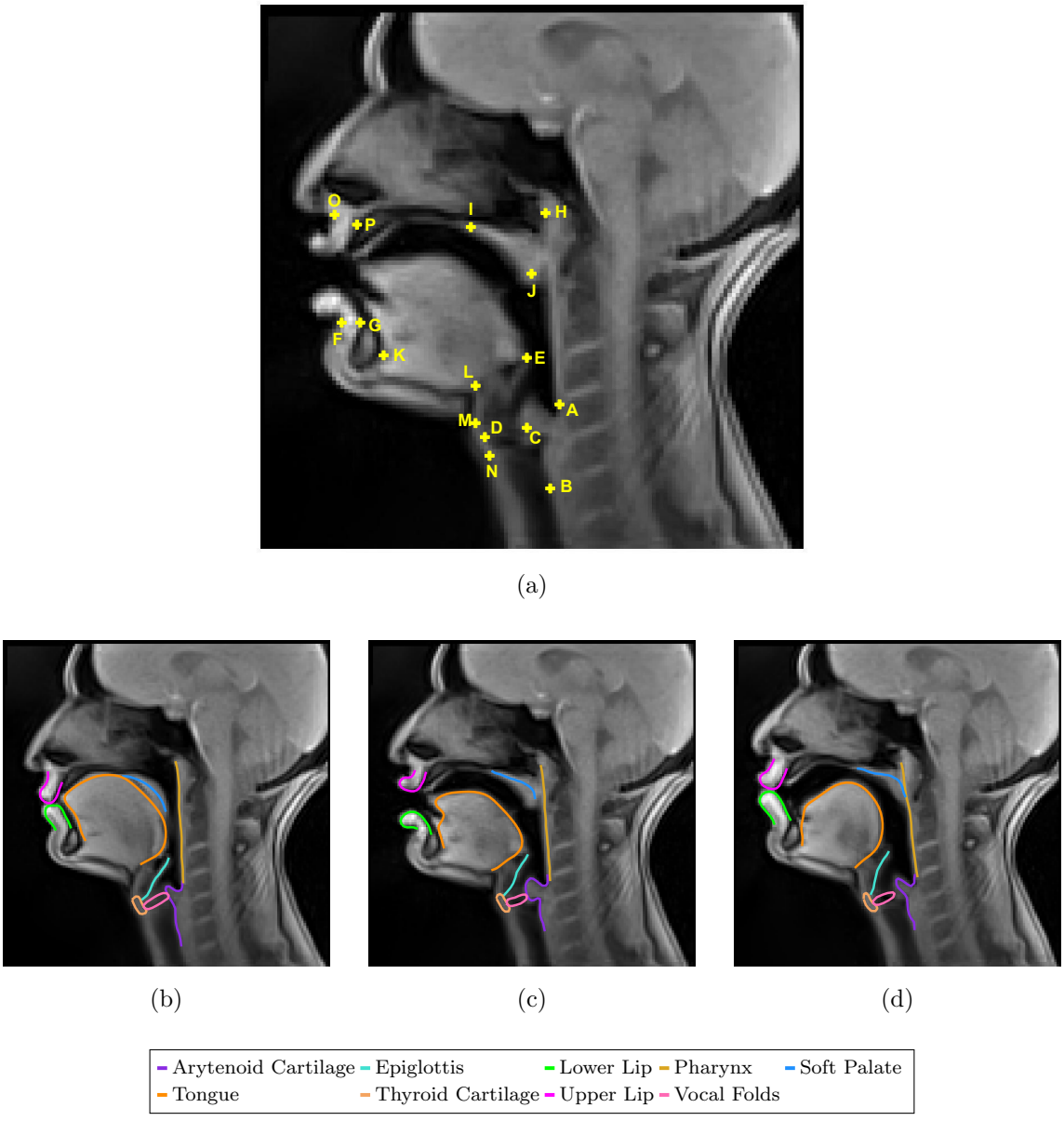


Figure 5.1: Landmarks in a mid-sagittal MRI sample and annotation samples exemplifying the procedure. (a) The mid-sagittal MRI sample shows the landmarks that were used to guide the annotation procedure described in Section 3.2. (b-d) The annotation samples show how the articulators were annotated in the MRI sample.

localizing the region of interest before segmenting it, which avoids spurious predictions in image regions that do not correspond to the articulator.

The models were pre-trained on the `COCO train2017` dataset [152], which contains RGB images of common objects. However, the MRI frames are grayscale. Therefore, we used the temporal dimension to build a more contextualized input. The network's inputs were formed by putting frames $t-1$, t, and $t+1$ in the first (R), second (G), and third (B) input channels, respectively.

4.2 Post-processing Algorithms

The post-processing of the network's outputs depends on the articulator. For each articulator, a specific algorithm is chosen, including additional sub-steps and adjustments of several hyper-parameters. This section explains the two approaches developed according to the articulator contour's closed or open nature.

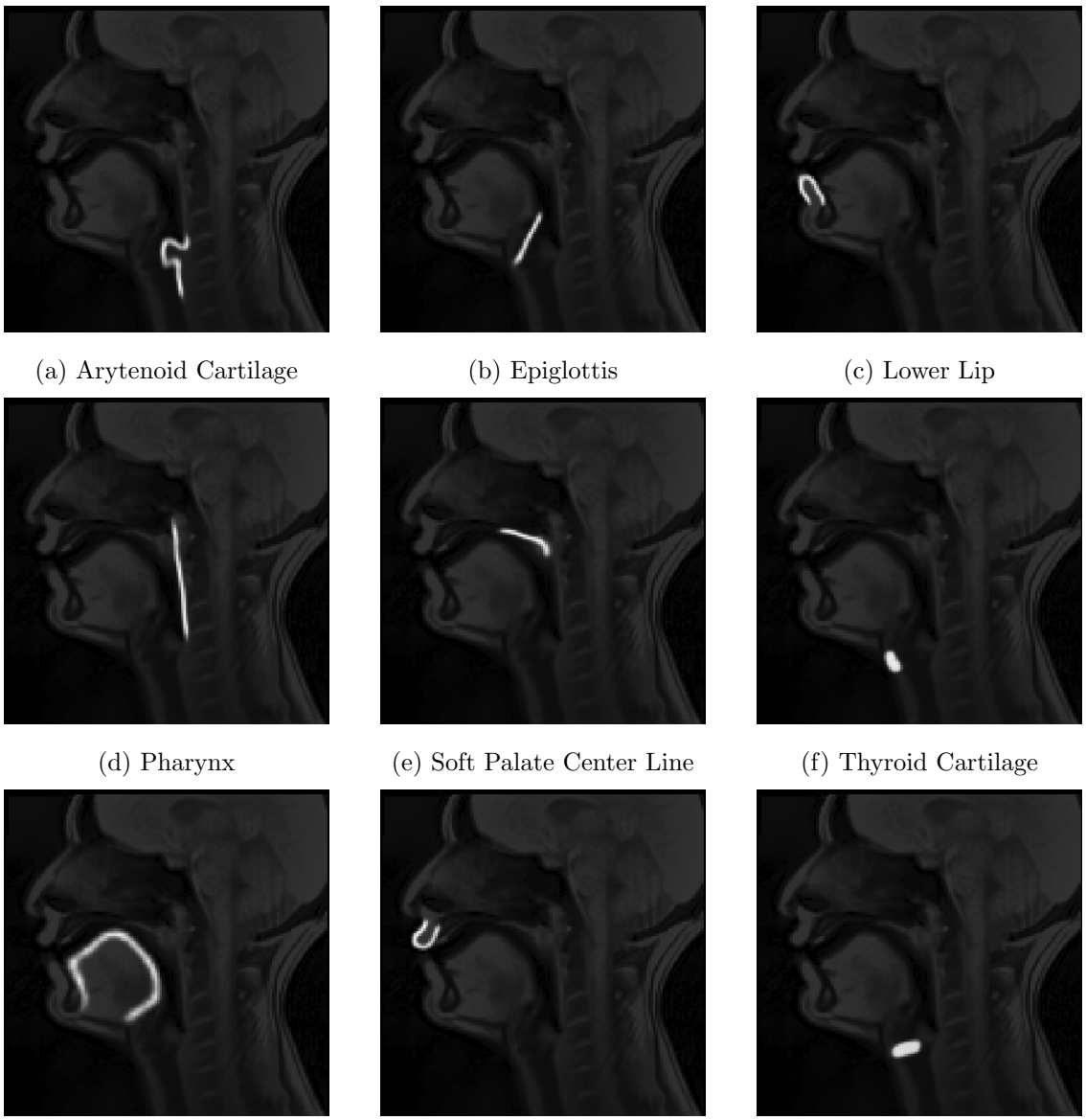
For articulators that were annotated as closed contours, we utilize the largest contiguous ISO-valued contour (Section 4.2). For open contours, we utilize a graph-based algorithm (Section 4.2). Figure 5.2 presents one segmentation mask sample for each articulator, illustrating the inputs of the post-processing algorithms.

Largest Contiguous ISO-valued Contour

For articulators annotated as closed contours, the contour can be found by calculating the largest contiguous ISO-valued contour in the probability map. We used the `find_contours` function[4] from `scikit-image` [291], which uses the marching squares method to compute the ISO-valued contours of the input 2D array for a particular level value. In our case, the level value is 1, obtained after thresholding the probability map. Figure 5.3 presents each step of the algorithm on a custom synthetic image.

We use this method for the **thyroid cartilage** and **vocal folds**, with thresholds of 0.7 and 0.8 for the pixel probability, respectively. In rare cases, the network may output two separate blobs for one articulator, producing two non-contiguous contours. In these cases, we choose to keep the largest area as the true contour.

[4] https://scikit-image.org/docs/0.8.0/api/skimage.measure.find_contours.html

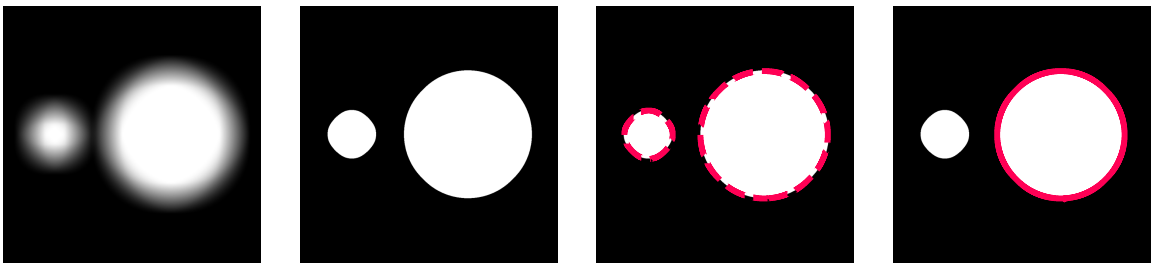


(a) Arytenoid Cartilage (b) Epiglottis (c) Lower Lip

(d) Pharynx (e) Soft Palate Center Line (f) Thyroid Cartilage

(g) Tongue (h) Upper Lip (i) Vocal Folds

Figure 5.2: Illustration of the segmentation mask for each articulator in one MRI sample. The segmentation masks are superimposed on the MRI sample with very low transparency to help the reader localize the articulator in the MRI. These segmentation masks are the inputs of the post-processing algorithms.



(a) Probability map input with multiple blobs

(b) Probability map after thresholding

(c) ISO-valued contours for both blobs

(d) Contour with the largest area

Figure 5.3: Steps of the largest contiguous ISO-valued contour algorithm for an illustrative artificial input.

Graph-based Algorithm

The open contours can be found by expressing the non-zero pixels in the segmentation mask as graph nodes and connecting the extremities using Dijkstra's shortest path algorithm [292]. We use this algorithm for the **arytenoid cartilage**, **epiglottis center line**, **lower and upper lips**, **pharynx**, **soft palate center line**, and **tongue**. The method requires a particular set of steps to be performed, which depends on the articulator.

The first step is thresholding to limit the number of nodes in the graph, which is done in two ways. In the first, the pixel value is given by

$$p_{\texttt{new}} = \begin{cases} 0, & \text{if } p_{\texttt{orig}} \leq T \\ 1, & \text{otherwise} \end{cases}$$

In the second, the pixel value is given by

$$p_{\texttt{new}} = \begin{cases} 0, & \text{if } p_{\texttt{orig}} \leq T \\ p_{\texttt{orig}}, & \text{otherwise} \end{cases}$$

where $p_{\texttt{orig}}$ is the original pixel value, $p_{\texttt{new}}$ is the updated pixel value, and T is the threshold value.

The second step is skeletonization performed using the `scikit-image`'s `skeletonize` function[5]. Then, the centers of the non-zero pixels in the image are converted to the nodes of a graph, and the edges between the nodes are created based on the Euclidean distance between

[5] https://scikit-image.org/docs/stable/api/skimage.morphology.html#skimage.morphology.skeletonize

Table 5.3: Steps and parameters of the graph-based algorithm for all articulators.

Articulator	Threshold Value (Type)	Skeletonize	Extremities Choice	Ang. Distance Reference	α	β
Arytenoid Cartilage	0.2 (0/1)	Yes	Angular distance	CM's y-coordinate + Right-most x-coordinate	1	10
Epiglottis	0.3 (0/1)	Yes	Vertical extremities	–	1	10
Lower Lip	0.4 (0/1)	Yes	Angular distance	CM	1	10
Pharynx	0.3 (0/1)	Yes	Vertical extremities	–	1	10
Soft Palate Center Line	0.1 (0/1)	Yes	Horizontal extremities	–	1	10
Tongue	0.2 ($0/p_{\texttt{orig}}$)	No	Angular distance	CM	10^{-7}	1
Upper Lip	0.4 (0/1)	Yes	Angular distance	CM	1	10

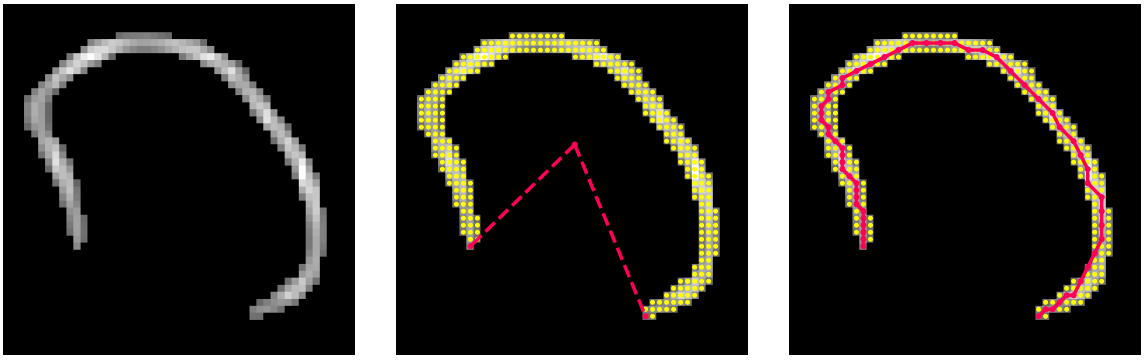
them and the probability of each node. The edge weight is given by

$$w_{ij} = \alpha \cdot d(i,j) + \beta \cdot (1 - p_j)$$

where $d(i,j)$ is the Euclidean distance between node i and node j and p_j is the probability of node j. An edge is set between two nodes if the infinity norm between them is lower than two pixels.

The next step is determining the contour extremities using one of three methods: the greatest angular distance, vertical, or horizontal extremities. The graph's center of mass (CM) is used as the reference for the greatest angular distance in all cases except for the arytenoid cartilage. For the arytenoid cartilage, only the CM's y-coordinate is used, and the x-coordinate is set to the right-most edge of the image. The two points with the greatest angular distance from the reference are selected as the extremities.

For the vertical extremities, the top-most and the bottom-most nodes in the graph are used, while for the horizontal extremities, the left-most and the right-most nodes in the graph are used. Finally, Diijkstra's algorithm is used to connect the two extremities, and the final contour is output. Table 5.3 summarizes the specific steps and parameters of the graph-based algorithm for each articulator. Figure 5.4 illustrates the algorithm's main steps for the tongue.



(a) Tongue's probability map after thresholding

(b) Graph nodes and contour's extremities given by the maximum angular distance

(c) Dijkstra's shortest path

Figure 5.4: Steps of the graph-based algorithm for an illustrative artificial input.

Summary

In summary, the main aspects of the proposed method are:

- **Inputs:** The input is the concatenation of MRI frames $t-1$, t, and $t+1$ forming an RGB input image;
- **Mask R-CNN**: The segmentation network used in the first phase. The network outputs a bounding box, a segmentation mask and a class probability for each articulator;
- **Post-processing**: The post-processing converts the segmentation masks produced by the Mask R-CNN into a curve describing the articulators' geometry. The largest ISO-valued contour was used for articulators with closed contours and the graph-based approach was used for the others.
- **Outputs:** The output of the system is the geometry of each articulator given by a vector $y_i \in \mathbb{R}^{N_{\texttt{samples}} \times 2}, \forall i \in \{1, ..., N_{\texttt{art}}\}$, where $N_{\texttt{samples}}$ is the number of samples in the curve and $N_{\texttt{art}}$ is the number of articulators. In our case, $N_{\texttt{samples}} = 50$ and $N_{\texttt{art}} = 9$.

4.3 Experimental Design

We aimed to develop a speaker-independent method to accurately and individually track non-rigid vocal tract articulators in RT-MRI movies. We also wanted to investigate how speaker adaptation could improve the method's performance for a new subject.

We carried out two experiments. The first was a LOOCV protocol. We removed subject S7.1/S7.2 from the test phase in the LOOCV pipeline because they account for the most images in the database, but still kept it for training. Leaving S7.1 and S7.2 out would have resulted

Table 5.4: Hyperparameters of the articulator segmentation model. This table summarizes the hyperparameters used to train and evaluate the segmentation network. The second part of the table refers to the speaker adaptation experiments.

Hyperparameter	Value
Batch size	8
Early-Stopping Patience	20
Weight Decay	10^{-3}
Sched. Max. Learning Rate	10^{-4}
Sched. Base Learning Rate	2×10^{-6}
Total Number of Network Parameters	44 401 393
Adapt. Num. Epochs.	20
Adapt. Learning Rate.	10^{-5}
Adapt. Sched. Red. Factor	10
Adapt. Sched. Patience	10

in a significant reduction in the training set, making it difficult to determine the cause of the performance improvement. From the remaining six subjects, we isolated one at a time and trained a model with the remaining subjects. We then tested the model on the test set of the left-out subjects. We also tested all of the LOOCV models on S7.1 and S7.2 test sets.

In the second experiment, we fine-tuned each initially trained model with its respective left-out subject. We did this using 10, 40, and all the training samples. We then evaluated the improvement in performance on the test sets of the left-out subject and the test sets of subjects S7.1 and S7.2. Ideally, the adapted model would improve its performance for the target subject while keeping the performance of the previously seen subjects constant.

The models were trained using the Adam optimizer [293] with the cyclic learning rate scheduling policy [108]. The training continued for the speaker adaptation experiments using the reduce learning rate on plateau scheduling policy. The hyperparameters of the training are given in Table 5.4. The machine learning code was developed using `PyTorch` [294]. The complete code for reproducing our results and using our software is available in our public repositories[6][7].

4.4 Evaluation

The performance of our models was evaluated using two main metrics: the root mean square (RMS) value of the point-to-closest-point distance (`P2CP`) described in Chapter 4 and the Jaccard index. The `P2CP` is unsuitable for closed curves, so the Jaccard index was also used. The

[6] https://github.com/vribeiro1/vocal-tract-seg
[7] https://gitlab.inria.fr/multispeech/vt/vt_tracker

Jaccard index, also known as Intersection-over-Union (IoU), is calculated by finding the intersection and union of the areas delimited by the target and the predicted contours. The Jaccard index ranges from 0 to 1, with 1 indicating a perfect match.

The inter-subject reproducibility of the results was assessed with the one-way ANOVA test for subjects S1-S6 and the unpaired t-test for subjects S7.1 and S7.2. $p < 0.05$ was considered significant.

5 Results

Generally, the proposed method demonstrated a good segmentation quality. Typical examples of the predicted and ground truth contours are shown in Figure 5.5 and Figure 5.6. They are representative of the overall method's performance, showing that it is adequate for different speakers and vocal tract positions. Figure 5.6 shows cases of swallowing (non-speech), which led to the worst results. It should be noted that we excluded swallowing images from the test set, and these cases are only discussed for completeness. Table 5.5 shows the P2CP_{RMS} values for each of illustrative case in millimeters.

Table 5.6 and Table 5.7 show the results in the LOOCV. Figure 5.7 and Figure 5.9a present the results in the form of boxplots to facilitate results visualization and comparison. The statistical test (one-way ANOVA) shows that the results were significantly different for all articulators except the soft palate center line. However, a visual analysis of Figure 5.7 and Figure 5.9a demonstrates that the low p-values are usually explained by a single outlier.

Table 5.8, Figure 5.8, and Figure 5.9b show the results when the models are evaluated on the S7.1 and S7.2 test sets. The statistical test demonstrates a significant difference between the two sets for all articulators except the lower lip and the soft palate center line. However, the differences between the mean P2CP_{RMS} distances tend to remain less than 0.5 mm for most articulators. The articulators that do not satisfy this condition are epiglottis, thyroid cartilage, and vocal folds.

The LOOCV results show that the overall performance has an error of less than 2.2 mm, slightly above one pixel (1.36 pixel). The segmentation of articulators annotated as closed contours (thyroid cartilage and vocal folds) provide a Jaccard index of about 60% in both cases. As pointed out, the tables and figures do not include swallowing cases.

Figure 5.10 and Figure 5.11 show the impact of speaker adaptation on model performance, and the x-axis represents the size of the adaptation training set; the y-axis represents the P2CP_{RMS}/Jaccard index value on the test set. The colored lines represent the results for each test subject, the pink (and blue) solid lines represent the mean metric, and the shaded regions

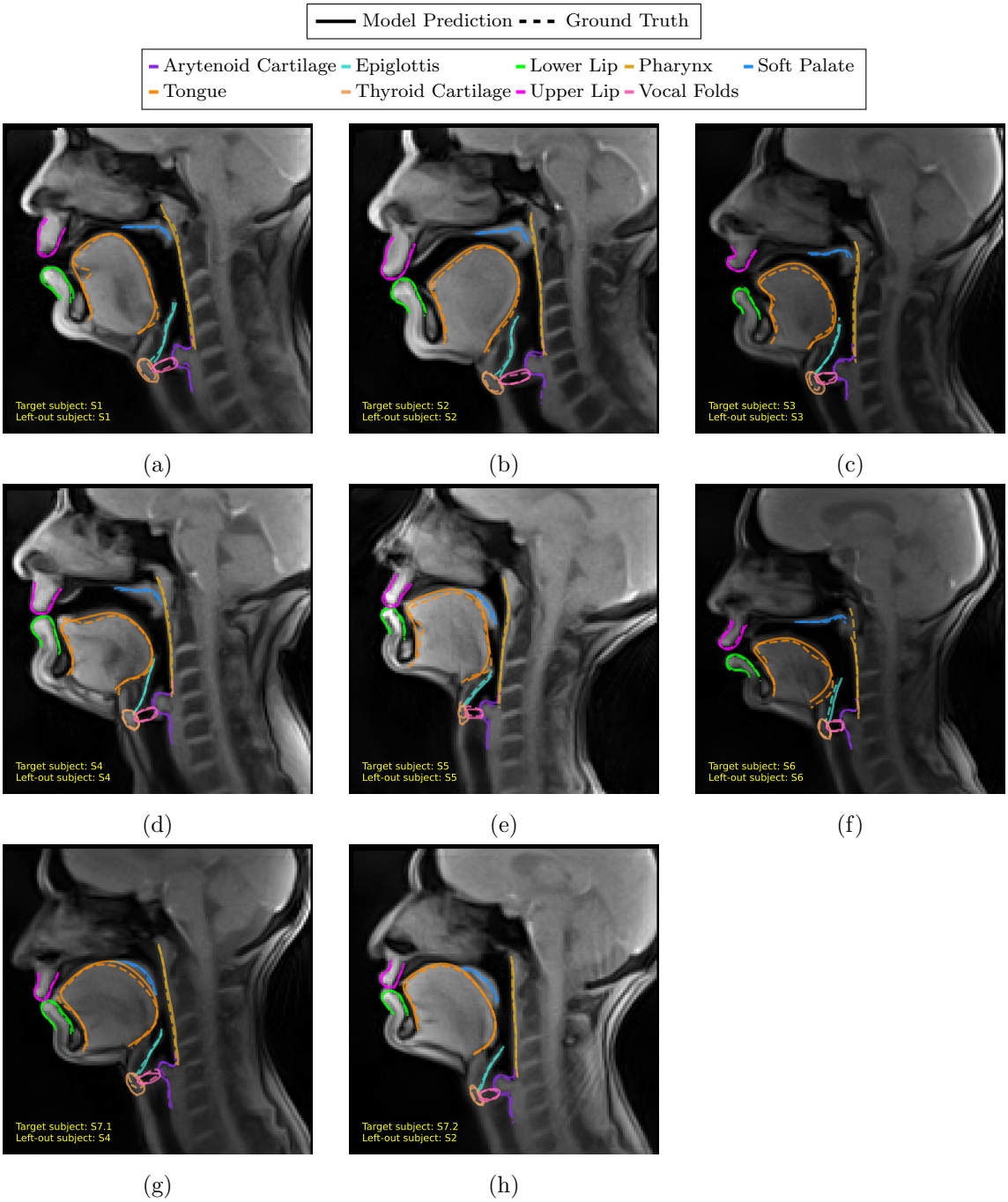
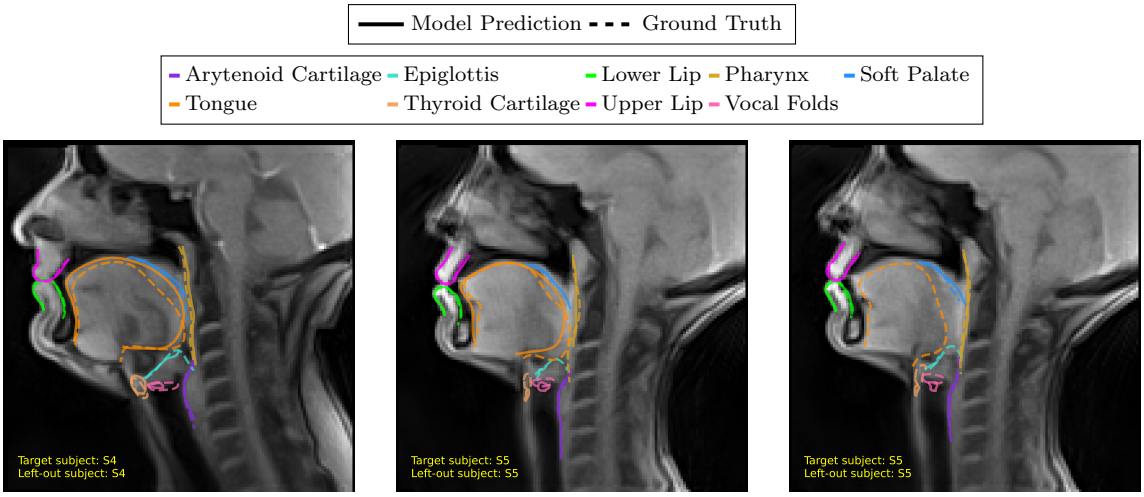


Figure 5.5: MRI samples of each subject superimposed with the predicted and ground truth contours after b-spline regularization. The text in the images indicates the ID of the subject in the image (target subject) and the ID of the left-out subject during the training of the model that produced that output. This figure shows how the predicted contours compare to the ground truth contours for each subject. The left-out subject is the subject that was not used to train the model.



(a) (b) (c)

Figure 5.6: MRI samples of swallowing superimposed with the predicted and ground truth contours after b-spline regularization. The text in the images indicates the ID of the subject in the image (target subject) and the ID of the left-out subject during the training of the model that produced that output. The figure shows how the predicted contours compare to the ground truth contours for each subject. The left-out subject is the subject that was not used to train the model. Note that for (c), the model completely missed the tongue.

represent one standard deviation. It can be seen that in case of initially poor inter-subject prediction, the $\texttt{P2CP}_{\texttt{RMS}}$ curve rapidly decreases after retraining with a small amount of additional images (usually ten).

Table 5.5: P2CP$_{RMS}$ error (in millimeters) and the Jaccard index (for closed contours) for the samples presented in Figure 5.5 and Figure 5.6.

Figure	Arytenoid Cartilage	Epiglottis	Lower Lip	Pharynx	Soft Palate Center line	Thyroid Cartilage		Tongue	Upper Lip	Vocal Folds	
	P2CP$_{RMS}$	P2CP$_{RMS}$	P2CP$_{RMS}$	P2CP$_{RMS}$	P2CP$_{RMS}$	P2CP$_{RMS}$	Jacc.Ind.	P2CP$_{RMS}$	P2CP$_{RMS}$	P2CP$_{RMS}$	Jacc.Ind.
Figure 5.6a	1.25	2.80	1.02	1.54	3.47	1.52	0.59	3.71	0.53	2.71	0.30
Figure 5.6b	5.23	5.14	0.95	4.82	1.55	5.28	0.06	3.75	0.93	2.32	0.36
Figure 5.6c	2.72	4.92	0.85	1.51	1.69	5.83	0.02		0.91	3.26	0.17
Figure B.2a	0.96	1.41	1.39	1.42	0.88	1.44	0.73	2.62	0.98	1.04	0.69
Figure B.2b	2.07	1.35	0.84	0.96	1.03	1.10	0.73	1.95	0.69	1.06	0.78
Figure B.2c	1.29	1.07	0.67	1.29	1.07	2.21	0.52	2.26	0.96	1.05	0.75
Figure 5.5d	1.12	0.52	0.69	0.91	1.23	0.63	0.89	1.79	0.64	0.72	0.85
Figure 5.5e	1.06	1.72	0.88	0.87	1.21	1.86	0.48	2.89	1.31	1.25	0.53
Figure 5.5f	0.56	2.22	1.18	3.82	0.99	0.59	0.85	2.33	1.20	0.57	0.81
Figure 5.5g	1.01	1.15	0.83	1.03	1.30	2.38	0.52	2.01	0.64	1.17	0.73
Figure 5.5h	1.25	0.80	0.53	0.73	0.73	0.87	0.70	1.63	0.77	0.77	0.73

Table 5.6: $\texttt{P2CP}_{\texttt{RMS}}$ error (in millimeters) for each articulator for each left-out subject in the LOOCV setting. The symbol † indicates cases in which the model missed that articulator in one or more images. The column in bold shows the mean ± standard deviation in each row. The p-values were calculated using a one-way ANOVA test using the subjects as the treatment variable.

Articulator	S1	S2	S3	S4	S5	S6	mean ± std	p-value
Arytenoid Cartilage	1.10 ± 0.41	2.08 ± 0.50	1.11 ± 0.17	1.17 ± 0.28	1.12 ± 0.30	0.92 ± 0.26	$\mathbf{1.25 \pm 0.42}$	1.80×10^{-19}
Epiglottis	1.15 ± 0.32	1.94 ± 0.83	0.90 ± 0.19	1.08 ± 0.33	1.56 ± 0.56	1.14 ± 0.62	$\mathbf{1.29 \pm 0.38}$	2.22×10^{-8}
Lower Lip	1.16 ± 0.26	0.73 ± 0.16	1.00 ± 0.24 †	0.86 ± 0.17	0.95 ± 0.22	1.03 ± 0.23	$\mathbf{0.96 \pm 0.15}$	2.01×10^{-7}
Pharynx	1.07 ± 0.28	0.98 ± 0.17	1.26 ± 1.35	1.07 ± 0.36	1.09 ± 0.17	2.28 ± 1.44	$\mathbf{1.29 \pm 0.49}$	9.34×10^{-6}
Soft Palate Center Line	1.01 ± 0.20	1.15 ± 0.30	1.14 ± 0.40	1.17 ± 0.47	0.98 ± 0.39	1.13 ± 0.45	$\mathbf{1.10 \pm 0.08}$	0.49
Thyroid Cartilage	1.70 ± 0.46	1.77 ± 0.46	1.75 ± 0.26	1.21 ± 0.42	3.78 ± 2.29	0.93 ± 0.27	$\mathbf{1.86 \pm 1.00}$	2.48×10^{-14}
Tongue	2.08 ± 0.42	1.91 ± 0.24	2.24 ± 0.42	1.93 ± 0.36	3.07 ± 0.67	1.87 ± 0.32	$\mathbf{2.18 \pm 0.46}$	8.62×10^{-16}
Upper Lip	0.84 ± 0.18	0.73 ± 0.12	0.81 ± 0.19	0.90 ± 0.39	1.20 ± 0.32	0.99 ± 0.25	$\mathbf{0.91 \pm 0.17}$	1.20×10^{-6}
Vocal Folds	1.35 ± 0.41	1.77 ± 1.21	1.20 ± 0.38	1.35 ± 0.45	2.84 ± 2.03	1.13 ± 0.65	$\mathbf{1.61 \pm 0.64}$	3.15×10^{-6}
mean ± std	$\mathbf{1.27 \pm 0.39}$	$\mathbf{1.45 \pm 0.55}$	$\mathbf{1.27 \pm 0.45}$	$\mathbf{1.19 \pm 0.31}$	$\mathbf{1.84 \pm 1.08}$	$\mathbf{1.27 \pm 0.48}$		0.2207

Table 5.7: Jaccard index for each closed contour for each left-out subject in the LOOCV setting. The column in bold shows the mean ± standard deviation in each row. The p-values were calculated using a one-way ANOVA test using the subjects as the treatment variable.

Articulator	S1	S2	S3	S4	S5	S6	mean ± std	p-value
Thyroid Cartilage	0.65 ± 0.10	0.61 ± 0.10	0.56 ± 0.07	0.70 ± 0.11	0.33 ± 0.19	0.74 ± 0.07	$\mathbf{0.60 \pm 0.14}$	2.61×10^{-20}
Vocal Folds	0.69 ± 0.09	0.59 ± 0.27	0.70 ± 0.09	0.67 ± 0.14	0.35 ± 0.25	0.67 ± 0.19	$\mathbf{0.61 \pm 0.13}$	5.65×10^{-8}

Table 5.8: The mean ± standard deviation of the $\texttt{P2CP}_{\texttt{RMS}}$ and Jaccard index (for closed contours only) when the models were tested with S7.1 and S7.2. The symbol † indicates cases in which the model missed that articulator in one or more images. The p-values were calculated using the unpaired t-test using the test subjects as the treatment variable.

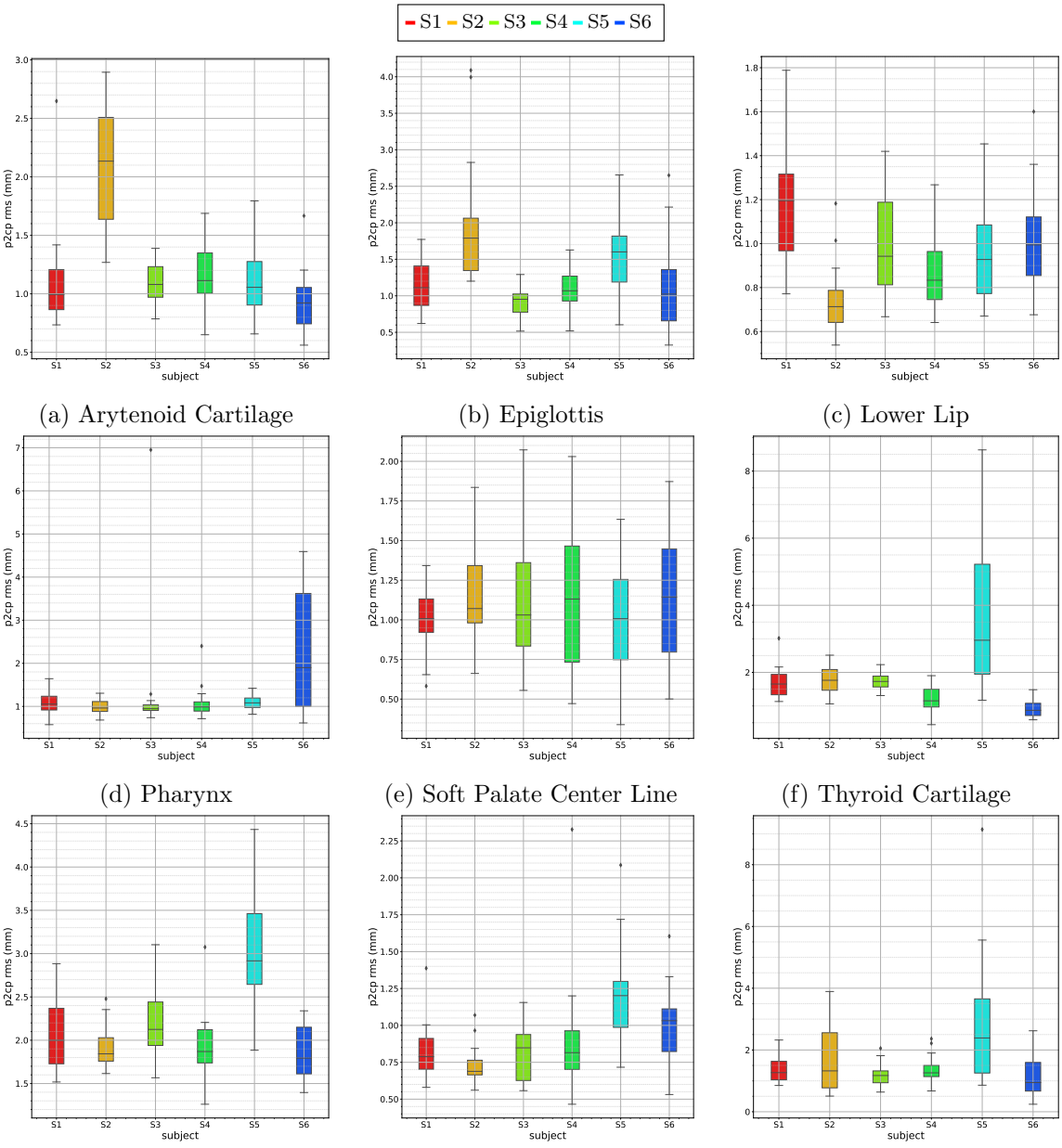
	$\texttt{P2CP}_{\texttt{RMS}}$ (mm)			Jaccard index		
Articulator	S7.1	S7.2	p-value	S7.1	S7.2	p-value
Arytenoid Cartilage	1.09 ± 0.03	1.20 ± 0.06	1.3×10^{-4}			
Epiglottis	1.53 ± 0.08	0.93 ± 0.08	10^{-64}			
Lower Lip	0.80 ± 0.03	0.79 ± 0.03 †	0.35			
Pharynx	0.84 ± 0.03	0.78 ± 0.02	10^{-3}			
Soft Palate	0.94 ± 0.02	0.96 ± 0.05	0.59			
Center line						
Thyroid Cartilage	2.09 ± 0.05	1.03 ± 0.03	10^{-158}	0.53 ± 0.01	0.69 ± 0.01	10^{-74}
Tongue	1.86 ± 0.06	1.39 ± 0.12	10^{-51}			
Upper Lip	0.86 ± 0.04	0.94 ± 0.02	1.8×10^{-3}			
Vocal Folds	1.64 ± 0.05	0.99 ± 0.03	10^{-65}	0.58 ± 0.01	0.74 ± 0.01	10^{-59}
mean ± std	$\mathbf{1.29 \pm 0.49}$	$\mathbf{1.00 \pm 0.19}$	0.1161			

6 Discussion

Our models can segment non-rigid vocal tract articulators with low error and outstanding generalization across subjects, as demonstrated by the LOOCV protocol. Although formal statistical analysis shows significant inter-subject variations of the mean annotation error, these differences are much less than the pixel size.

The model generally performs poorly for the sublingual cavity, as observed in Figure B.2a and Figure 5.5e. It happens because the shortest path algorithm can sometimes miss accentuated curvatures. Nevertheless, the acoustic relevance of the sublingual cavity is minor compared to the tongue tip and tongue dorsum.

Another notable case is Figure 5.5e, where the tongue deviation is close to 3 mm. In this case, the divergence is due to a possible inconsistency in the annotation decision. The annotator selected a more internal part of the tongue body, while the model delineated a more external contour. These discrepancies are usual in machine learning. The performance metrics are calculated in reference to a human annotator. However, even if different specialists provide their annotations for the same images, the annotations are likely to differ. This effect is known as inter-annotator agreement, a common phenomenon in image segmentation tasks [178]. A slight deviation from the ground truth is acceptable, but the hypothetical inter-annotator agreement should constrain it. Otherwise, the model would copy a specific specialist instead of learning the



(a) Arytenoid Cartilage (b) Epiglottis (c) Lower Lip

(d) Pharynx (e) Soft Palate Center Line (f) Thyroid Cartilage

(g) Tongue (h) Upper Lip (i) Vocal Folds

Figure 5.7: Distribution of the $\texttt{P2CP}_{\texttt{RMS}}$ error (in millimeters) for each articulator for each left-out subject in the LOOCV setting. Similar information is also shown in Table 5.6. Note the different y-scales when comparing the plots.

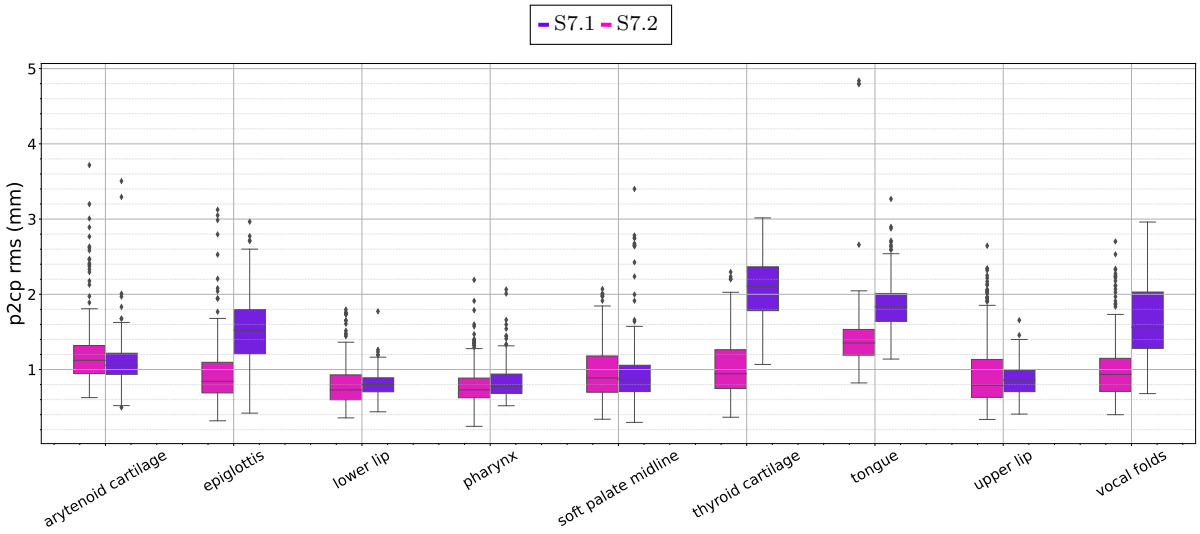
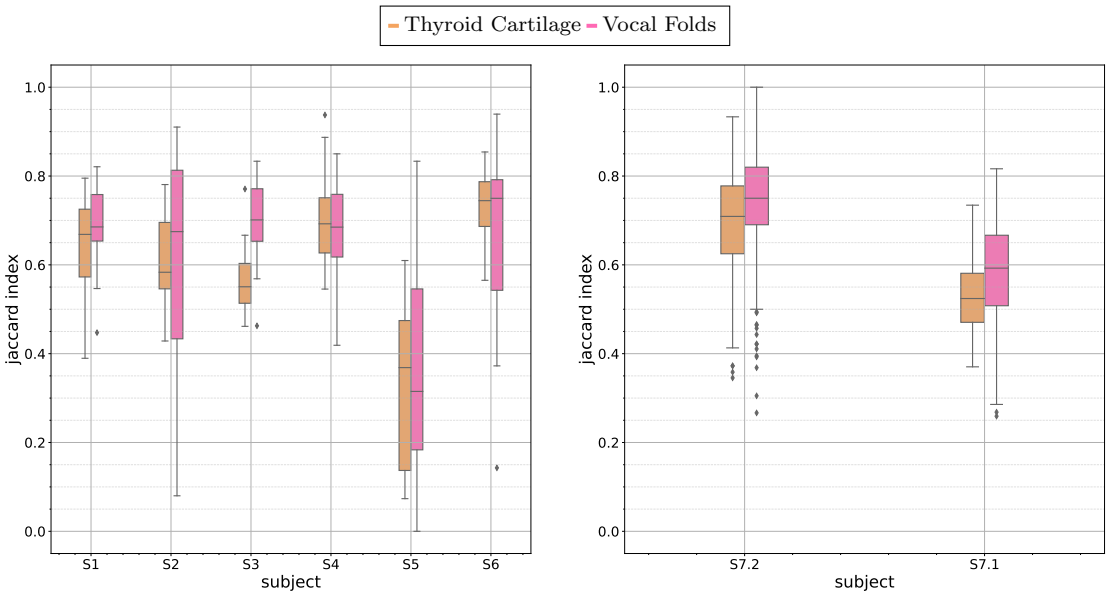
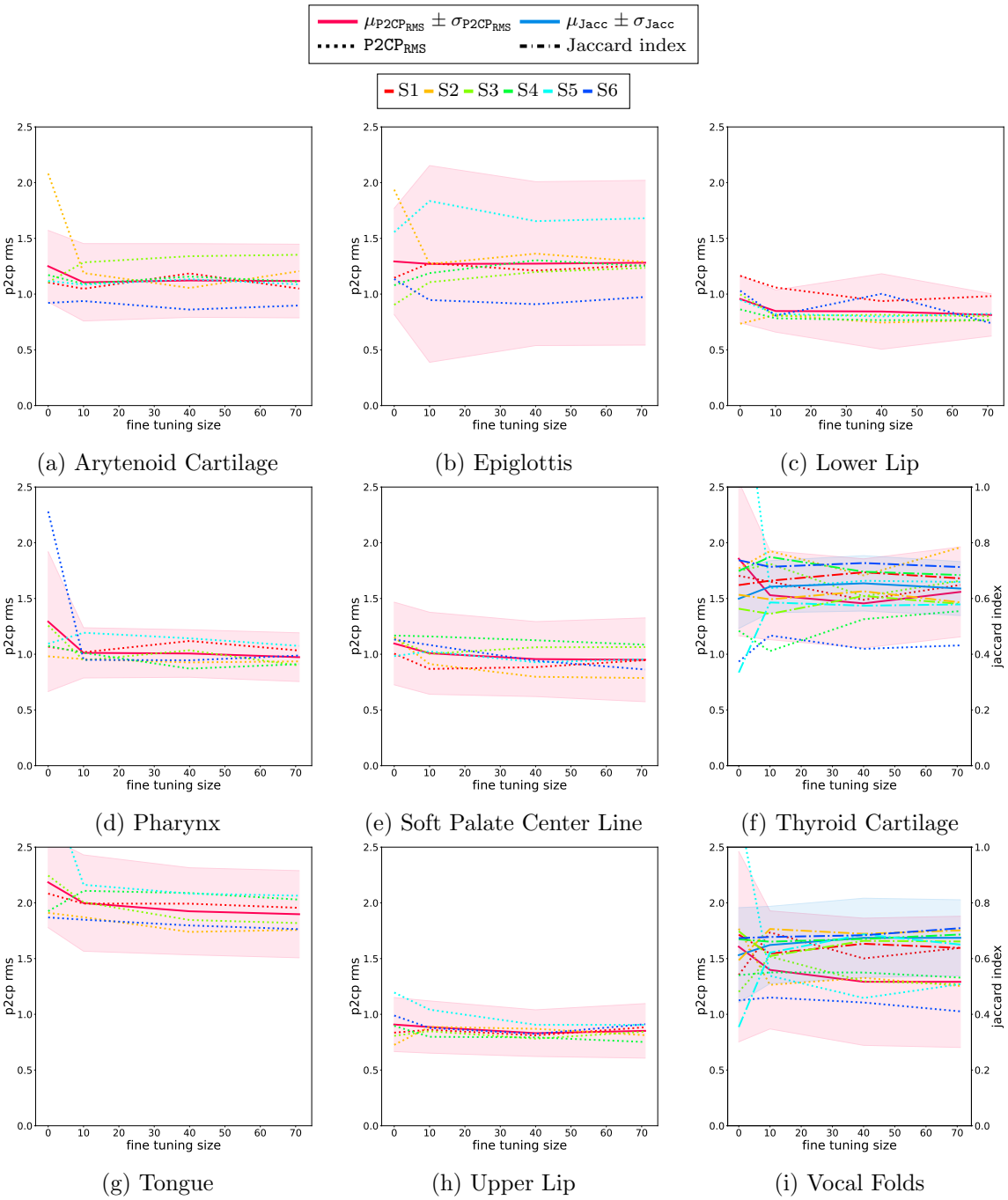


Figure 5.8: Distribution of the $\texttt{P2CP}_{\texttt{RMS}}$ error (in millimeters) for each articulator for S7.1 and S7.2. Similar information is also shown in Table 5.8. Note the y-scales when comparing with Figure 5.7.



(a) LOOCV (b) S7.1 & S7.2

Figure 5.9: Distribution of the Jaccard index for the articulators with closed contours (a) for each left-out subject and (b) for S7.1 and S7.2. Similar information is also shown in Table 5.8.

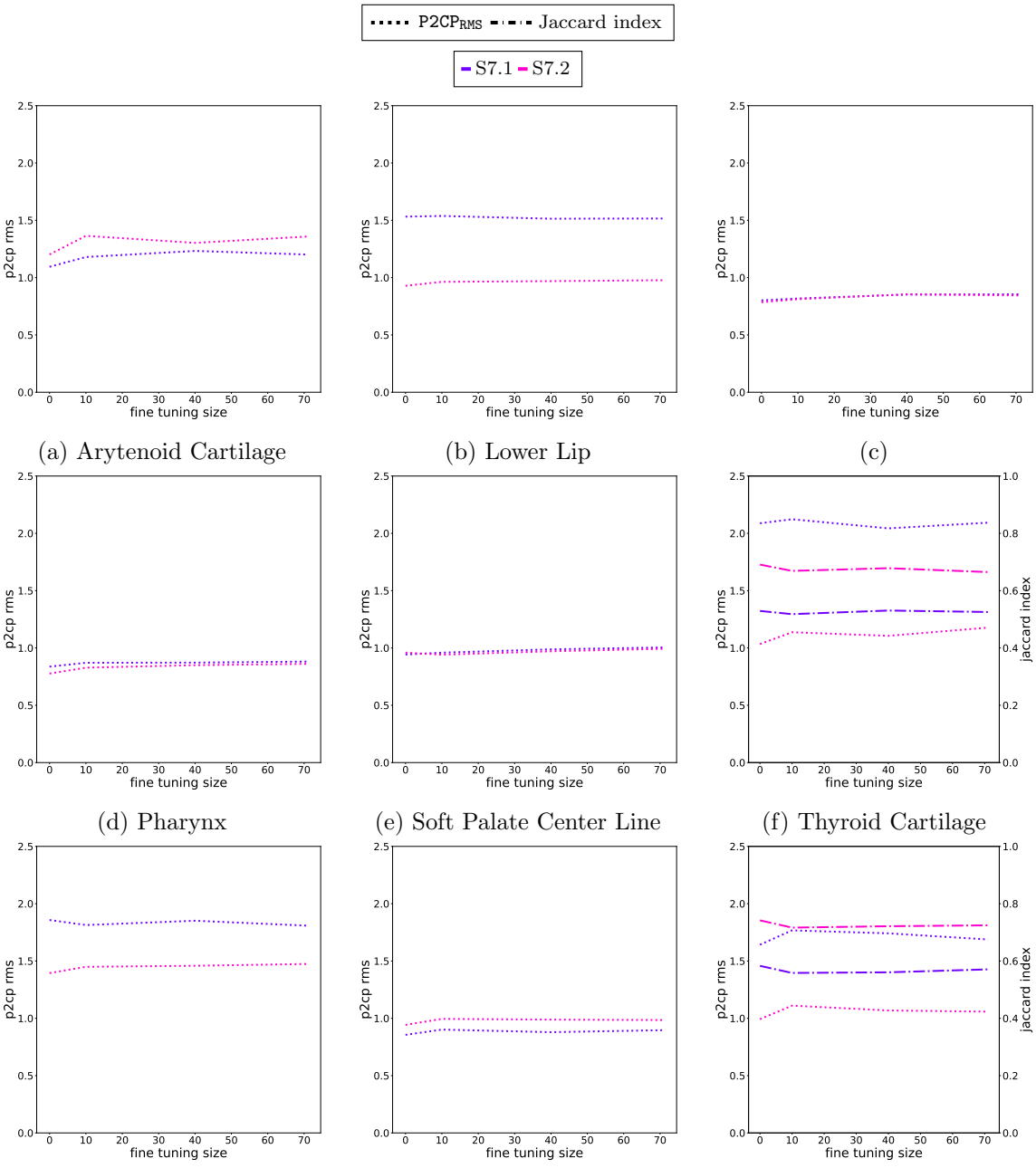


(a) Arytenoid Cartilage (b) Epiglottis (c) Lower Lip

(d) Pharynx (e) Soft Palate Center Line (f) Thyroid Cartilage

(g) Tongue (h) Upper Lip (i) Vocal Folds

Figure 5.10: P2CP$_{\text{RMS}}$ and Jaccard index (only for closed contours) calculated on the test sets of the left-out subjects for each articulator when the models were adapted with varying numbers of training samples of the left-out subject.



(a) Arytenoid Cartilage (b) Lower Lip (c)

(d) Pharynx (e) Soft Palate Center Line (f) Thyroid Cartilage

(g) Tongue (h) Upper Lip (i) Vocal Folds

Figure 5.11: P2CP$_{\text{RMS}}$ and Jaccard index (only for closed contours) calculated on the test sets of S7.1 and S7.2 for each articulator when the models were adapted with a varying number of training samples of the left-out subject.

task. However, an extended analysis of the vocal tract segmentation inter-annotator agreement is beyond the scope of this research.

The predictions for the laringeal articulators (arytenoid cartilage, epiglottis, thyroid cartilage, and vocal folds) are adequate, which is encouraging since this is a challenging region for human annotators. The most significant errors for the epiglottis occur at its extremities, as seen in Figure 5.5f. The prediction is correctly located for the thyroid cartilage and the vocal folds, and the errors are related to the size of the articulator; the algorithm usually yields a larger area. However, for articulatory speech research, the correct location of vocal folds is much more important than the precise contour since it is the source of the voice and directly impacts synthesis and control models.

For S5, the head is slightly rotated, as the subject seems to be leaning upwards, contrary to the others leaning forward. This case differs from a simple image rotation, which data augmentation could easily handle. The head rotation produces a slight deformation in the larynx, which is more pronounced in the thyroid cartilage and vocal folds region. For this reason, the model struggles to accurately predict their shapes when subject S5 is left out of the training set (Table 5.7). On the other hand, when the model is adapted to it, the performance improvement is very noticeable.

The case when the contour's extremities accounts for the largest errors also affects other articulators such as the tongue, pharynx, and soft palate. The contour's extremities also accounts for the largest errors for other articulators such as the tongue, pharynx, and soft palate. A few cases for the tongue are visible in Figure B.2c and Figure 5.5e. Nevertheless, the difference is less acoustically relevant for most of these articulators since they occur in a region that does not alter the final vocal tract air column.

Figure 5.6 shows a few cases of swallowing, which led to the worst results. Swallowing is an essential human process but is also one of the most difficult to annotate and predict. It is because the epiglottis lies over the arytenoid cartilage during swallowing, covering the glottis and preventing anything but air from entering the lungs. It creates constrictions between the articulators and the bolus, removing air-tissue boundaries. As a result, the articulators are almost indistinguishable, making the annotation difficult even for specialists.

Not surprisingly, the model provides the most unreliable results for swallowing cases. The tongue and the epiglottis errors are around 3 mm, and the Jaccard index for the vocal folds and thyroid cartilage is low. In some cases, the model even misses some articulators completely. It would have been possible to obtain better results for swallowing by significantly increasing the number of swallowing examples in the training set. However, this is not the focus of our

work, which is concerned with articulatory speech research. We leave improving the results for swallowing for future work.

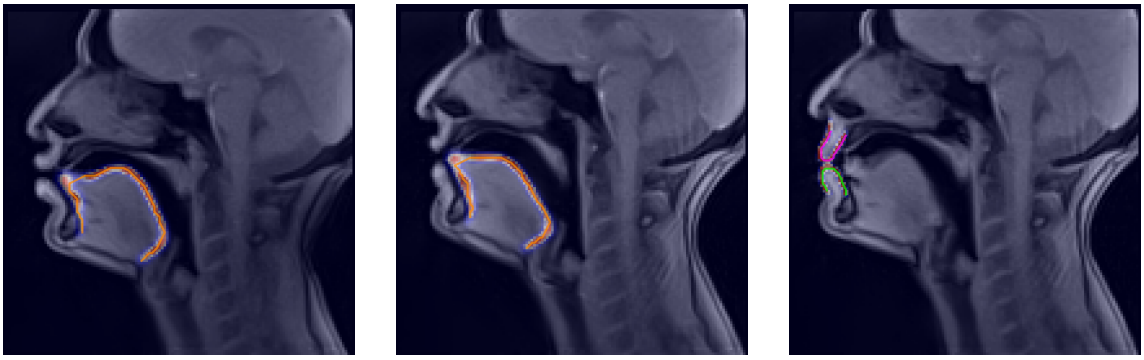
We compared our work to that of Labrunie et al. [266] due to the similarities we found between them. Labrunie et al. [266] used RT-MRI data and considered most of the articulators as we did, except for the thyroid cartilage and vocal folds. However, the two studies had substantial differences in the annotation decisions. Labrunie et al. [266] did not include the sublingual cavity, starting the tongue annotation from the tip, and they chose to annotate the contour of the epiglottis and soft palate, not only the center lines. Most importantly, the two studies used different test sets. Despite these differences, the similarities between the two studies allow for some level of comparison. For a fair benchmark, it is desirable to have access to the same images and a standardized annotation procedure.

On average, our results are close to the Multiple Linear Regression (MLR) while under performing compared to the best approach – the mASM. However, our method has the advantage of being speaker-independent by design, while the method proposed by Labrunie et al. [266] is subject-specific, limiting the impact on the community.

The second experiment evaluated the adaptation to an unseen speaker. The results from Figure 5.10 suggest that there is a significant improvement in the results with only ten additional training images when the model initially performed poorly, such as the vocal folds and thyroid cartilage for S5 and the pharynx for S6, even though on average the gain, if any, is minimal. The result indicates that adaptation is beneficial when the target subject has a more pronounced anatomical or postural differences from the training subjects. The adaptation gain is lower for cases where the target subject is standardized, such as the same head position.

During the speaker adaptation procedure, the model could specialize in the new subject and "forget" the previous ones. This phenomenon is known in the machine learning literature as catastrophic forgetting. It could be avoided using elastic weight consolidation [295]. However, the results of the second experiment shown in Figure 5.11 suggest that the model does not forget the previous subjects. It is likely because the speaker adaptation procedure was conservative, using a lower learning rate for a few epochs, preventing the model from diverging but also limiting the improvement.

It is essential to note that our method has limitations. The main one is not including rigid articulators, such as the jaw, upper incisor, and hard palate. These articulators are indispensable for speech production but are challenging to segment in MRI because bones are indistinguishable from the air in the image, so we can only observe the teeth root trace, which contains a small amount of water. Since these articulators are rigid, the problem is restricted



(a) (b) (c)

Figure 5.12: MRI samples that illustrate cases where the model failed to predict the contact between articulators. The images include the original MRI sample, the predicted segmentation mask, and the contour without b-split regularization.

to finding their location in the image. Therefore, an alternative for tracking these articulators would be sliding a pre-computed mask in the image and retrieving the image region with the highest correlation with the mask. However, it can be laborious because of the requirement for manual adjustments in the tracking.

Furthermore, it is difficult to track contacts between articulators, especially for the tongue tip, as evidenced by the samples in Figure 5.12, which shows MRI samples with a segmentation mask overlay and the shortest path contour without b-split regularization. Although the most extreme point in the tongue tip might have a higher probability of belonging to the contour, it corresponds to a much longer path, which the shortest path algorithm rejects; thus, the contact between the tongue tip and the alveolar region is missed.

Finally, another limitation concerns the RT-MRI frame rate. Our method was trained with a frame rate of 50 frames per second, meaning that each frame corresponds to 20 milliseconds of speech. However, the constriction interval associated with some phonemes, such as /l/, have a shorter duration than the MRI frame, which means the image will be blurred, and the true articulator position will be uncertain. In these cases, the performance of the model may be lower.

7 Conclusions

In this study, we proposed and assessed a method for segmenting the vocal tract shape in real-time MRI using a deep learning approach. We also proposed a transfer learning method operating with a small dataset (in comparison to other deep learning applications). The method accurately estimated the shapes of nine non-rigid articulators that delimit the vocal tract and are essential for articulatory speech synthesis. We also showed that the model can generalize

to new subjects. If the position of a new subject significantly differs from the previous set, the model may need to be adapted to the target subject. Nevertheless, we showed that only a small amount of data (about ten images) is required for a good adaptation.

Our recent research in articulatory synthesis of speech has shown that the method is helpful for this task [18, 20, 17]. We have developed a `Python` package that makes exploring RT-MRI data to investigate speech production easy. Finally, `ASD1` images are already available [87]. We plan to publicly release the `ASD2` dataset together with the manual and automatic annotations for both datasets.

We hope this work will be helpful for other researchers interested in investigating speech production using RT-MRI data.

Chapter 6

Automatic Synthesis of the Vocal Tract Shape

> Machines take me by surprise with great frequency.
>
> *Alan Turing*

1 Overview

Many challenges surround articulatory synthesis, particularly the non-uniqueness problem, speaker normalization, and critical articulation. Non-uniqueness means many different vocal tract configurations may result in the same spectral characteristics. Speaker normalization is related to the anatomical differences between speakers, which require the normalization of the vocal tract to generate relevant shapes for any speaker. The difficulty is to separate the variability related to speech strategy vs. vocal tract anatomy. Articulatory phonology [234] defines the speaker's gestures and gesture scores to produce each phoneme. In this context, an articulator is critical for a particular phoneme if a specific place of articulation is required for its utterance. If an articulator is not critical, it is said to be free for that phoneme.

The importance of these issues is that they pose fundamental challenges to synthesizing a phonetically relevant vocal tract shape. For example, suppose the synthesized shape for /p/ lacks lip closure; in this case, the phonetic relevance is lost even if many other metrics are within the expected boundaries. Conversely, when comparing two different speakers, point-wise measurements are likely to fail even if the correct places of articulation are satisfied.

This Chapter describes our approach to speech articulation synthesis and evaluation. Since all experiments are based on the same dataset, with the same training, validation, and test

splits, we dedicate Section 2 to describe the data and the splitting strategy.

Section 3 and Section 4 describe our approach to vocal tract shape synthesis from the sequence of phonemes to be articulated. The first proposes a recurrent neural network (RNN) that directly predicts each vocal tract articulator individually from the sequence of phonemes to be articulated, and the second proposes a method that uses an autoencoder-based articulatory model as an auxiliary articulatory model to the main phoneme-to-articulation RNN. Section 5 focuses on the problem of evaluating synthetic vocal tract shapes. We propose a phoneme recognition task to measure the phonetic information retained by the mid-sagittal contours and reconstructed by the synthesizers presented in Section 3 and Section 4. Section 6 gives our final remarks in the Chapter.

A substantial part of this work was presented in Ribeiro et al. [18], Ribeiro et al. [20], and Ribeiro and Laprie [17], but using different datasets in most of them. The code to reproduce our experiments is available at our repository[1].

2 Dataset

2.1 Data Description

We used the dataset referred to as `ArtSpeech Database 2` (`ASD2`) in Chapter 5. Section 3.1 from Chapter 5 details the acquisition protocol, parameters, and recording conditions. In this Chapter, we use the entire dataset available at the time of writing. It contains the data of one female native French speaker articulating 1 629 utterances in French. The dataset accounts for 2 hours and 27 minutes of speech, with 439 018 MRI frames after excluding non-speech intervals. The MRI and audio recordings were collected in the laboratory IADI at the Centre Hospitalier Régional Universitaire de Nancy, France, and the participant provided written informed consent.

The exact time intervals between phones and speech gestures is critical for our work; thus, alignment errors substantially impact the results. Therefore, the phonetic alignment was obtained using forced alignment and then manually corrected by an expert in phonetics (the thesis supervisor). The phonetic vocabulary comprises 50 tokens, from which 42 are phonetic and eight are non-phonetic, representing unknown, blank, silence tokens, and noises after /i, e, u, y, ø/at the end of the sentence. Plosives are characterized by two phases: a closure and a burst. Therefore, the phonemes /p, b, t, k/ [2] are represented by one token for each phase. Table 6.1 details the vocabulary.

[1] https://github.com/vribeiro1/artspeech

[2] /b/ was split into closure and burst for completeness since the separation is less relevant for this phoneme.

Table 6.1: Description of the vocabulary used in the articulation synthesis experiments. Tokens are sorted by their order of appearance in the vocabulary.

Token	IPA Symbol	Meaning
BLANK		Blank token
UNK		Unknown token
#		Silence
2	ø	ø
2h		Noise after ø
9	œ	œ
@	ə	ə
E	ɛ	ɛ
E/	ɛ/e	Intermediate sound between ɛ and e
H	ɥ	ɥ
J	ɲ	ɲ
N	ŋ	ŋ
O	ɔ	ɔ
O/	o/ɔ	Intermediate sound between o and ɔ
R	r	r
S	ʃ	ʃ
U~/	ɛ̃/œ̃	Intermediate sound between ɛ̃ and œ̃
Z	ʒ	ʒ
a	a	a
a~	ã	ã
b	b	Burst of b
b_cl	b	Closure of b
d	d	Burst of d
d_cl	d	Closure of d
e	e	e
eh		Noise after e
f	f	f
g	g	g
i	i	i
ih		Noise after i
j	j	j
k	k	Burst of k
k_cl	k	Closure of k
l	l	l
m	m	m
n	n	n
o	o	o
o~	õ	õ
p	p	Burst of p
p_cl	p	Closure of p
s	s	s
t	t	Burst of t
t_cl	t	Closure of t
u	i	u
uh		Noise after u
v	v	v
w	w	w
y	y	y
yh		Noise after y
z	z	z

The model from Chapter 5 was used to track the articulators' contours for all images in the database. We manually corrected the few cases where the model missed some articulators. Data cleaning and correction are laborious, requiring many hours of validation and manual annotation. Therefore, we restricted the efforts to the cases that concern our research. Since non-speech articulations (silence and swallowing) are out of the scope of this thesis, we did not correct any of these cases.

2.2 Upper and Lower Incisors

In Chapter 2, we discussed the challenges of using MRI data. One of the most significant problems is the difficulty of tracking rigid structures, such as bones, due to the low T2*. Nevertheless, the hard palate line, the upper incisor, and the mandible are of utmost importance for the completeness of the vocal tract modeling for a few reasons. First, model training requires a

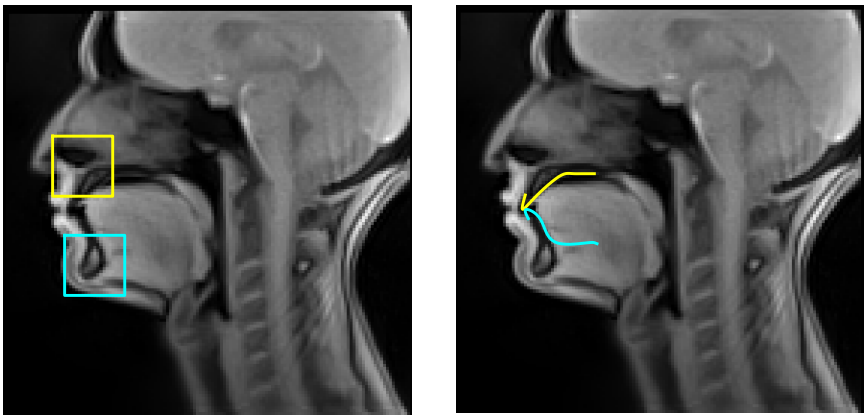
coordinate system with a fixed reference. The reference should be located at a rigid structure in the patient's head to correct for head movement and be equally sensitive to translations as the rest of the vocal tract. The upper incisor matches these constraints, being a good reference.

Second, constrictions between the tongue tip and the alveolar region are crucial for articulating labio-dental phonemes. Likewise, constrictions between the tongue dorsum and the mouth ceiling are necessary for articulating palatal phonemes. Third, the jaw is the primary articulation since it corresponds to the mouth opening. In addition, the relationship between mouth opening and tongue elevation is relevant for studying compensatory aspects of speech. Nevertheless, the mandible is not fully visible in the mid-sagittal plane, and the lower incisor's root is the single visible part. Due to its oval appearance in the MRI films, the exact jaw position is uncertain since this articulator is capable of angular, lateral, and protrusive movements.

The tracking system presented in Chapter 5 does not include the upper and lower incisors, the mandible, and the hard palate line. Since these articulators are rigid, we can assume a fixed shape, limiting the task of locating them in the MRI. The upper incisor and the hard palate, referred to simply as the upper incisor, were merged. Likewise, the lower incisor and the jawline, referred to simply as the lower incisor, were combined. Figure 6.1b illustrates the case for one MRI frame.

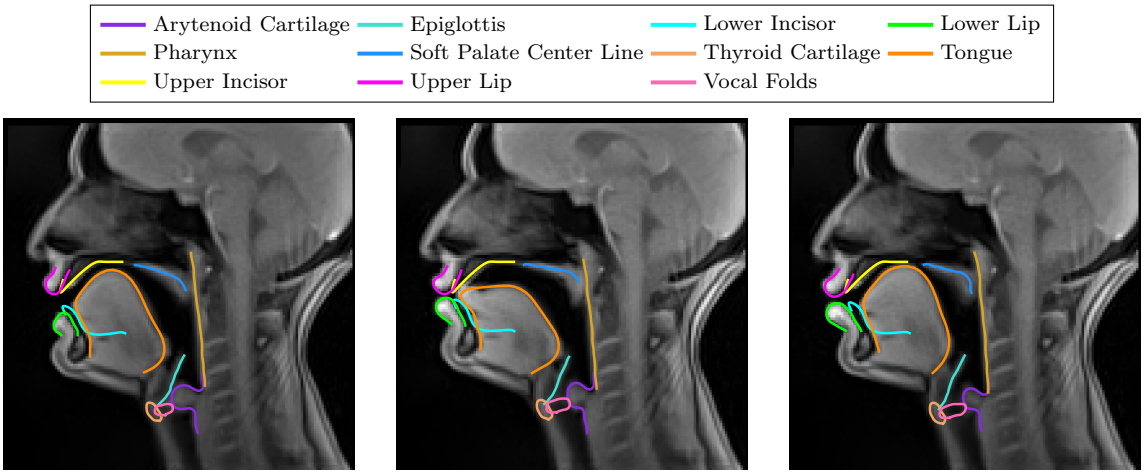
A region of interest (RoI) is annotated for a single frame (Figure 6.1a) per subject to localize these structures. Then, the RoI is slid through the target image for the subsequent frames, and the structural similarity index is computed [296]; the position with maximal similarity is selected, and the rigid body is drawn on that location. A search window in the image is predefined to reduce the space for locating each RoI. During data acquisition, the subject's head was fixed; therefore, we assume that any rotation in the upper part of the skull is negligible. However, the same is not valid for the mandible. To account for the jaw angular displacement, we generate rotated versions of the lower incisor's RoI and compute the similarity metric with these augmented masks. We select the position and angle pair that maximizes the similarity. For each frame, we allow a maximum variation of 2° in the angle from the previous frame.

With this method, we finalize the tracking of the complete vocal tract shape, from the glottis to the lips, for use in the articulatory synthesis of speech. The advantage of this procedure is that it does not demand training and requires a single weak annotation per speaker. Since the upper incisor procedure depends only on the current frame, it is very efficient and amenable to parallel processing of different frames. However, it this is not true for the lower incisor. Since we restrict the angular variation between subsequent frames, we must process the films in order.



(a) (b)

Figure 6.1: (a) RoIs for each rigid articulator and (b) lower incisor representing the jaw line and upper incisor representing the alveolar region plus the hard palate line.



(a) (b) (c)

Figure 6.2: MRI samples with superposed articulators.

2.3 Further Considerations

Figure 6.2 presents three dataset samples illustrating all articulators. Due to the impossibility of double-checking and correcting all of the automatic annotations present in the final dataset, it is important to stress that the data might contain annotation errors, which should be in the order of magnitude presented in the results section of Chapter 5. These errors occur mainly when two articulators are in contact and can compromise the quality of the vocal tract synthesis.

The data were randomly split into train, validation, and test sets. Table 6.2 presents the overall statistics for each split. As in Chapter 5, the splits were done at the level of acquisitions (80 seconds per acquisition) since they might contain repetitions of the same utterance with similar articulatory manners.

Table 6.2: Summary of the train, validation and test splits for articulatory synthesis.

	Number of Utterances	Number of MRI frames	Duration (minutes)
Train	1 399	372 004	125.1
Validation	116	33 615	11.3
Test	114	33 399	11.2
Total	1 629	439 018	147.6

3 Model-Free Vocal Tract Shape Synthesis from the Sequence of Phonemes to be Articulated

Synthesizing the vocal tract shape during speech requires modeling the interaction of several articulators while accounting for the context surrounding each phoneme production. Typically, coarticulation takes two forms: backward, due to inertia, and forward, due to anticipation. Nevertheless, as described by the Cohen and Massaro [222], the coarticulatory dominance of a phoneme exponentially decays with time. Therefore, the short-term dependencies in the phonetic sequence are expected to be much more relevant than long-term ones.

From the neural network designs assessed in Chapter 3, recurrent architectures are the most appropriate to model the problem due to their simplicity and lower data needs. Therefore, in this Section, we designed a speech articulation synthesizer for a single speaker based on an RNN capable of estimating the vocal tract shape for a sequence of phonemes to be articulated. We reference this method as model-free, meaning that it does not rely on an articulatory model of the vocal tract in opposition to our next approach. We compare our results with a simple baseline based on the average vocal tract shape per phoneme, which processes phonemes independently from the context.

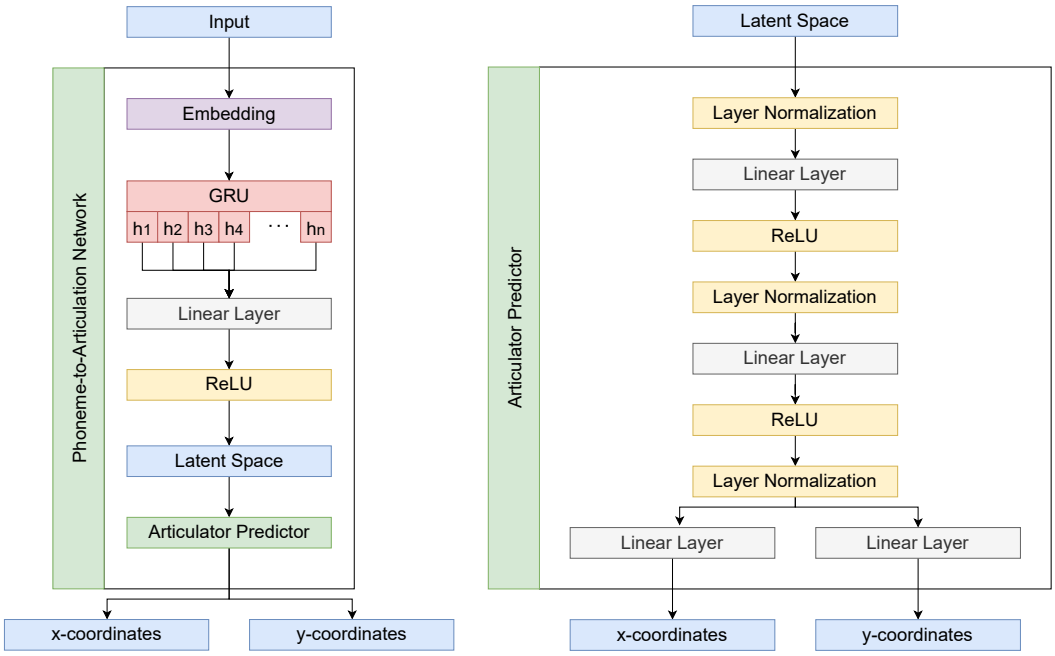
The designed approach produces realistic articulations, outperforming the baseline in all considered metrics.

3.1 Methods

Phoneme-Wise Mean Contour

It is not easy to find a baseline to compare our work with because few studies have used the same data and protocols as we have. Most research uses EMA data [241]. Csapó [243] used MRI but targeted the entire frame instead of individualized contours. Therefore, we proposed a simple method called phoneme-wise mean contour as a baseline inspired by concatenative speech synthesis.

The phoneme-wise mean contour works in two phases. In the first phase, it constructs a



(a) Phoneme-to-Articulation Network (b) Articulator Predictor

Figure 6.3: Model-Free Phoneme-to-Articulation network.

lookup table that maps each occurrence of a phoneme to its respective contour. The second phase takes an input sequence and searches the lookup table for each phoneme. The average shape for each articulator is then computed. It is essential to notice that this method does not involve learning or optimizing weights, as the lookup table only contains pre-computed data. Even though it is straightforward, this method does not consider the surrounding phonemes, so it does not account for coarticulation.

Model-Free Phoneme-to-Articulation

The proposed method is based on the architecture presented in Figure 6.3. The input of the network is the sequence of phonemes to be articulated with the phone interval duration provided by forced alignment encoded as token repetitions such that each token in the input sequence has the exact duration of its corresponding frame in the MRI sequence, i.e., with a 50 fps, each MRI frame corresponds to 20 ms of speech; hence, a phoneme with t ms is repeated $\frac{t}{20}$ times in the input sequence.

The encoder network contains two layers of bidirectional GRU (BiGRU). The hidden state at each time step inputs a linear layer with `ReLU` activation, forming the network's latent space. The decoder network is referred to as the Articulator Predictor Block. It contains one block per articulator, which only shares the latent space, i.e., the Articulator Predictor blocks do not share weights. The Articulator Predictors are composed of a sequence of layer normalization [207]

Table 6.3: Hyperparameters of the model-free phoneme-to-articulation training. The second part of table refers to the neural network hyperparameters.

Hyperparameter	Value
$N_{\texttt{art}}$	10
$N_{\texttt{samples}}$	50
Batch size	12
Early-Stopping Patience	30
Weight Decay	10^{-5}
Learning Rate (LR)	10^{-4}
LR Sched. Patience	10
Sched. Reduction Factor	10
Embedding Dim.	64
RNN Hidden Size	128
RNN Dropout	0.1
Predictor Hidden size	256
Total Number of Network Parameters	1 739 496

and linear layers with `ReLU` activation and output the x- and y-coordinates for each sample in the articulator curve at each time step.

The models were trained to minimize the Euclidean distance between the predicted and the target curves. Even though each Articulator Predictor block makes predictions independently, they are trained jointly. The loss function is given by

$$\mathcal{L}(p, \hat{p}) = \frac{1}{T \times N_{\texttt{art}} \times N_{\texttt{samples}}} \sum_{t=1}^{T} \sum_{i=1}^{N_{\texttt{art}}} \sum_{j=1}^{N_{\texttt{samples}}} d(p_{t,i,j}, \hat{p}_{t,i,j})$$

where T is the sequence length obtained after encoding the phoneme duration, p and $\hat{p}$ are the ground truth and the predicted curves, respectively, $N_{\texttt{art}}$ is the number of articulators, $N_{\texttt{samples}}$ is the number of samples in each curve, and d is the Euclidean distance. The models were trained with the Adam optimizer [293] with the reduce learning rate on plateau scheduling policy [3]. Our hyperparameters choices are detailed in Table 6.3.

In summary, the main aspects of the proposed method are:

- **Inputs:** The input to the network is a sequence of phonemes represented by its index in a vocabulary ($x \in \mathbb{R}^T$). The phoneme duration is encoded by repeating the token in the input sequence to match the number of MRI frames corresponding to that phoneme.

- **Bidirectional GRU:** As explained in Chapter 3, a bidirectional GRU is a recurrent neural network that processes information in both directions. It allows the model to learn

[3] https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html

contextual dependencies in the input sequence.

- **Latent space:** The latent space is a high-dimensional representation of the input sequence. This representation is internally shared between all Articulator Predictor blocks.
- **Articulator Predictor Block:** The architecture detailed in Figure 6.3 is used to generate the articulator curves. The blocks only share the input latent space.
- **Outputs:** The network outputs are the contours of each articulator at each time step. The final output is $y \in \mathbb{R}^{T \times N_{\texttt{art}} \times 2 \times N_{\texttt{samples}}}$.

Evaluation Strategy

The evaluation uses two main metrics: the reconstruction error measured in terms of the $\texttt{P2CP}_{\texttt{mean}}$ and the correlation coefficients between four tract variable trajectories – both metrics were detailed in Chapter 4. We are most concerned with the TVs LA, TTCD, TBCD, and VEL (see Table 4.1). The measurements were done in millimeters to facilitate result interpretation.

3.2 Results

Table 6.4 and Figure 6.4 present the $\texttt{P2CP}_{\texttt{mean}}$ for each articulator. The first presents the mean and standard deviations, while the latter presents the distributions as box plots to facilitate comparison. The same presentation is done for Pearson's correlation for the four TVs, depicted in Table 6.5 and Figure 6.5.

To permit better comprehension of the results, Figure 6.6 and Figure 6.7 present the TV trajectories and the corresponding tokens for two utterances in the test set. The tokens can be mapped to the phonemes using Table 6.1. Figure 6.6 presents each TV's trajectories produced by the model-free phoneme-to-articulation and ground truth TV trajectories. Additional samples are available in Section 1 of Appendix A. Furthermore, Figure 6.7 presents the ground truth, the baseline, and the trajectories of the proposed method for another utterance in the test set. The green arrows in Figure 6.6 indicate cases in which an articulatory target was reached, and the red arrows indicate the opposite.

3.3 Discussion

The two evaluation strategies (individual contours and TVs) show that the proposed model outperforms the baseline by a large margin for all articulators and all tract variables. The baseline system is a simplistic method that only considers the average articulation of each phoneme, missing vital information regarding coarticulation. For that reason, we observe monotonous TV

Table 6.4: Reconstruction error ($\texttt{P2CP}_{\texttt{mean}}$) for the phoneme-wise mean contour and the model-free phoneme-to-articulation. The best results are marked in bold.

Articulator	Method	$\texttt{P2CP}_{\texttt{mean}}$
Arytenoid Cartilage	Phon.-Wise Mean Contour	2.20 ± 1.00
	Model-Free Phon.-to-Art.	$\mathbf{1.77 \pm 0.86}$
Epiglottis Center Line	Phon.-Wise Mean Contour	2.12 ± 1.39
	Model-Free Phon.-to-Art.	$\mathbf{1.55 \pm 1.10}$
Lower Incisor	Phon.-Wise Mean Contour	1.81 ± 1.02
	Model-Free Phon.-to-Art.	$\mathbf{1.46 \pm 0.78}$
Lower Lip	Phon.-Wise Mean Contour	1.82 ± 0.85
	Model-Free Phon.-to-Art.	$\mathbf{1.39 \pm 0.67}$
Pharynx	Phon.-Wise Mean Contour	1.14 ± 0.50
	Model-Free Phon.-to-Art.	$\mathbf{1.07 \pm 0.46}$
Soft Palate Center Line	Phon.-Wise Mean Contour	1.76 ± 1.00
	Model-Free Phon.-to-Art.	$\mathbf{1.48 \pm 0.84}$
Thyroid Cartilage	Phon.-Wise Mean Contour	1.74 ± 0.91
	Model-Free Phon.-to-Art.	$\mathbf{1.53 \pm 0.87}$
Tongue	Phon.-Wise Mean Contour	3.32 ± 1.37
	Model-Free Phon.-to-Art.	$\mathbf{2.19 \pm 0.88}$
Upper Lip	Phon.-Wise Mean Contour	1.03 ± 0.39
	Model-Free Phon.-to-Art.	$\mathbf{0.90 \pm 0.34}$
Vocal Folds	Phon.-Wise Mean Contour	2.45 ± 1.02
	Model-Free Phon.-to-Art.	$\mathbf{1.88 \pm 1.02}$

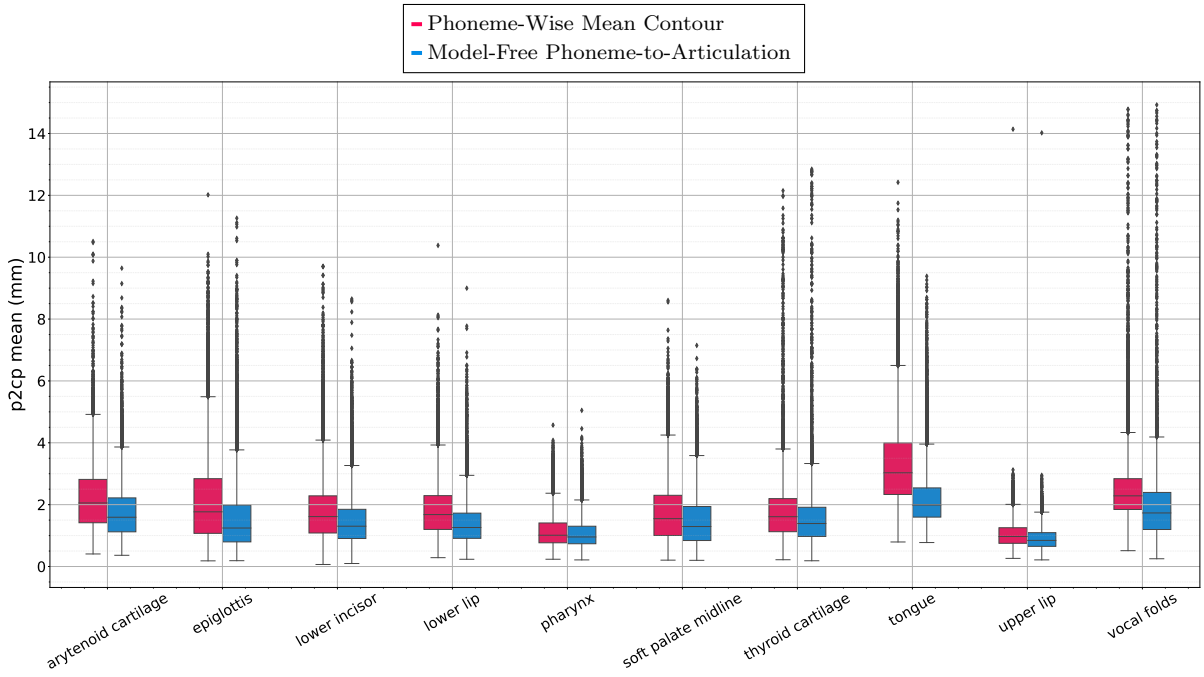


Figure 6.4: Box plots of the reconstruction error for the phoneme-wise mean contour and model-free phoneme-to-articulation for each articulator.

Table 6.5: Correlations between the target and predicted tract variables trajectories. The best results are marked in bold.

Tract Valiable	Method	Correlation	Min Correlation	Max correlation
LA	Phon.-Wise Mean Contour	0.46 ± 0.16	-0.40	0.73
	Model-Free Phon.-to-Art.	$\mathbf{0.80 \pm 0.16}$	-0.41	0.95
TTCD	Phon.-Wise Mean Contour	0.54 ± 0.13	0.01	0.80
	Model-Free Phon.-to-Art.	$\mathbf{0.86 \pm 0.07}$	0.66	0.97
TBCD	Phon.-Wise Mean Contour	0.51 ± 0.15	-0.17	0.74
	Model-Free Phon.-to-Art.	$\mathbf{0.82 \pm 0.08}$	0.44	0.94
VEL	Phon.-Wise Mean Contour	0.53 ± 0.12	0.18	0.78
	Model-Free Phon.-to-Art.	$\mathbf{0.74 \pm 0.18}$	-1.00	0.90

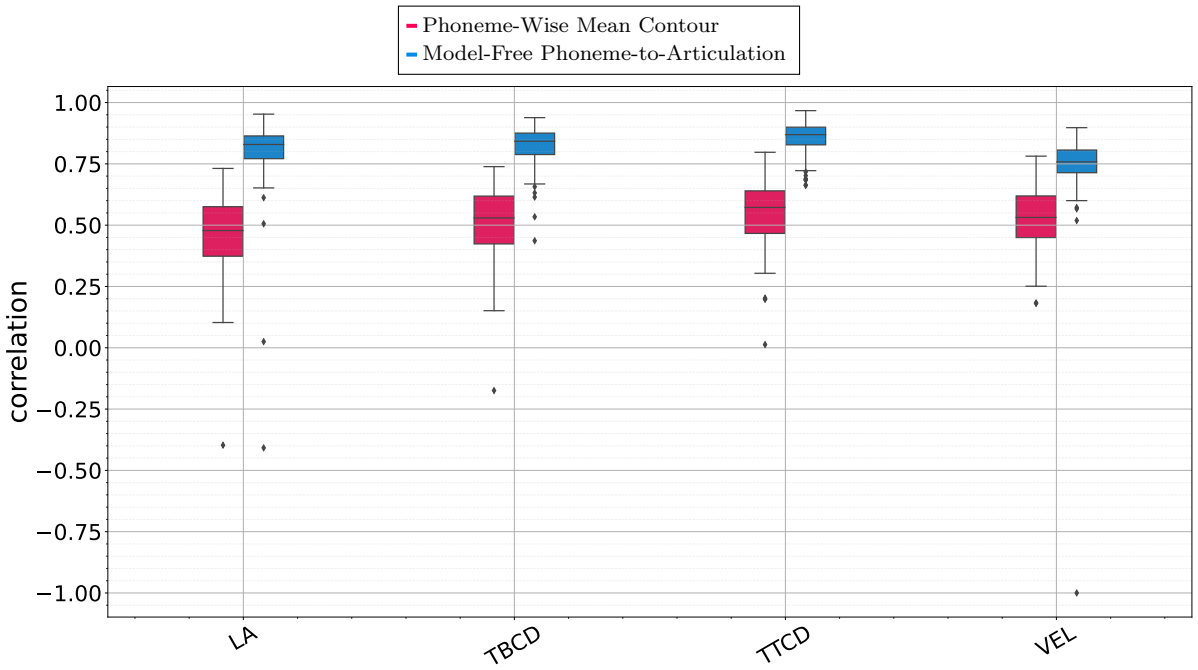
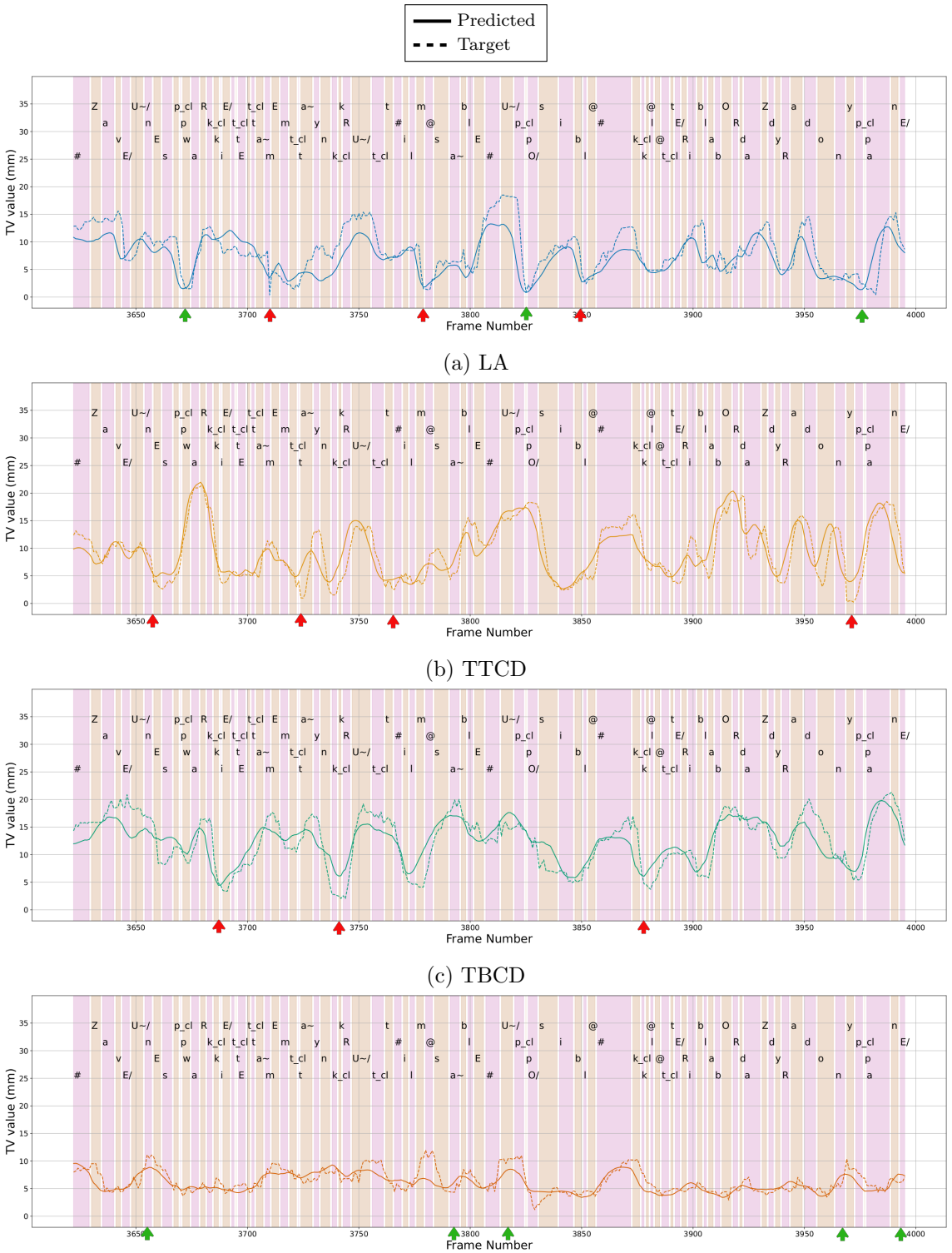


Figure 6.5: Box plots of the TVs correlations for the phoneme-wise mean contour and model-free phoneme-to-articulation.



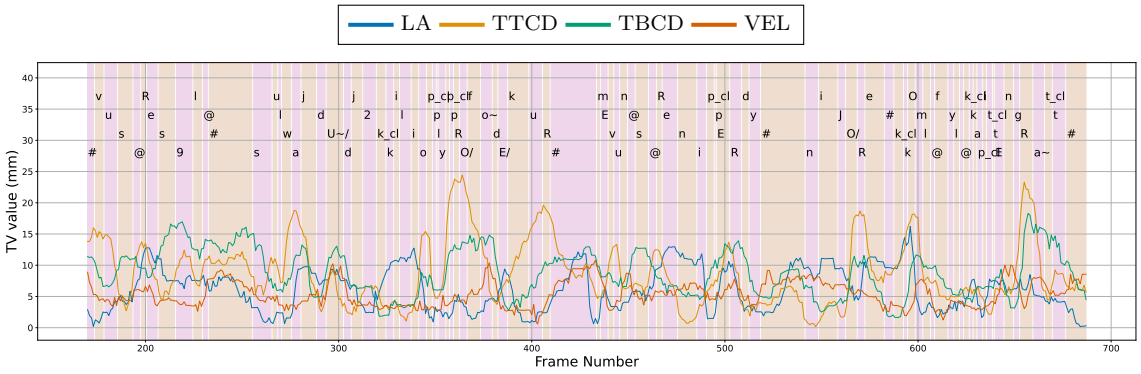
(a) LA

(b) TTCD

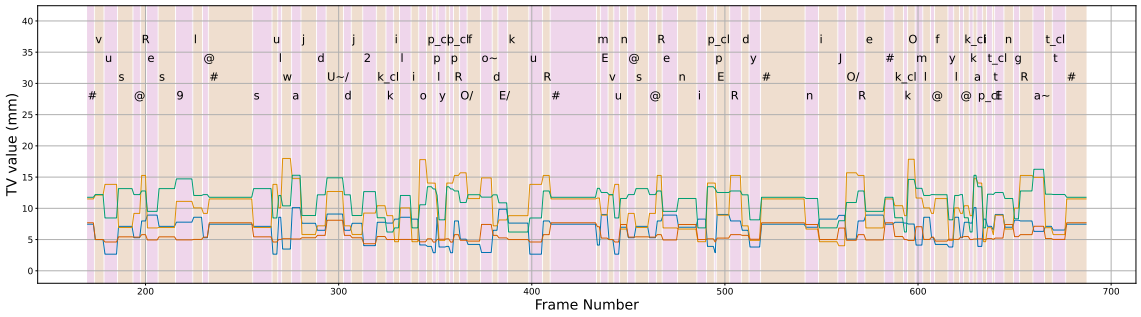
(c) TBCD

(d) VEL

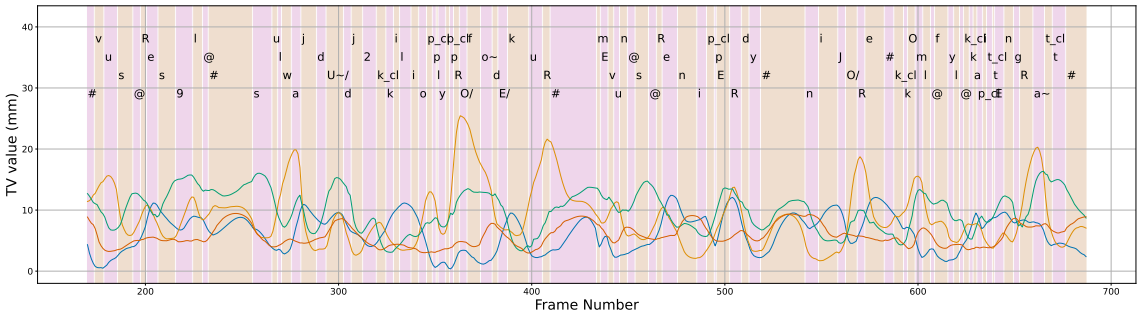
Figure 6.6: Model-free phoneme-to-articulation and ground truth TV trajectories for the utterance *"J'avais un espoir qui était en même temps une crainte, il me semblait impossible que le terrible abordage du radeau n'eût pas anéanti."* Each image displays one tract variable. The corresponding phonemes are displayed in the top of the image. Green and red arrows indicate whether the model reaches an articulatory target or not, respectively. The alternating colors mark the onset and offset of each phoneme.



(a) Ground Truth



(b) Phoneme-Wise Mean Contour



(c) Model-Free Phoneme-to-Articulation

Figure 6.7: Ground truth, phoneme-wise mean contour prediction, and model-free phoneme-to-articulation prediction for the utterance *"Vous serez seul, sous l'ouaill d'un Dieu qui lit au plus profond des cours, mais vous ne serez ni perdu, ni ignoré comme le fut le capitaine Grant."* The corresponding phonemes are displayed in the top of the image. The alternating colors mark the onset and offset of each phoneme.

trajectories, with significant steps between phonemes, looking unrealistic regarding the target curve. Better results would be expected by considering n-grams instead of single tokens in the computation of the average contour.

Conversely, the model-free phoneme-to-articulation can produce realistic vocal tract shapes that follow the expected trajectories correctly. From Table 6.5, we can observe that even in the worst executions of TTCD and TBCD, the model performs considerably well, with a reasonable correlation between the target and the predicted trajectories. The worst executions of LA and VEL are not good. However, it is essential to highlight that negative correlations occur only for one utterance for LA and one for VEL. The second worst execution of LA has a correlation of 0.03. Nevertheless, Figure 6.5 shows that these are outliers; when they are discarded, the worst execution of LA has a correlation of 0.51, and the worst execution of VEL has a correlation of 0.52. Both are at the same level as the other two TVs.

These results align with those from Ribeiro et al. [20], which showed that these negative correlations typically occur when the vocal tract variable is free in the execution of the entire phonetic sequence. Likewise, the current model presents more minor errors for the individual contours and higher correlations for the TV trajectories than the ones given in Ribeiro et al. [20]. The difference is probably linked to the datasets. The current dataset is more extensive than that of Ribeiro et al. [20], which has only 38 minutes of speech. Moreover, the present corpus has a broader phonetic context and longer utterances, contributing to phonetically rich synthesis.

Figure 6.6 and Figure 6.7 shows that for many cases where a TV is critical for a phoneme, the model-free phoneme-to-articulation yields vocal tract shapes with the expected dynamics, i.e., we observe the speech gestures such as lip closure (LA) for labials, tongue-dental closure (TTCD) for dentals, tongue-palatal closure (TBCD) for palatals, and velum opening (VEL) for nasals. However, the model tends to under-articulate, meaning it is rare to observe the critical places of articulation being fully reached. The arrows in Figure 6.6 and Figure 6.7 illustrate some of these cases. Interestingly, in many cases, the speaker also under-articulates, which is indicated by the dashed line. Thus, even though the model does not reach some articulatory targets, it copies the speaker's movements.

The problem of missing articulatory targets has two sources. The first is related to tracking errors that are especially prominent during contact. Tracking errors have a minimal impact when they occur in a region of a large cross-sectional area; however, they heavily affect the acoustics when they happen at a constricted point. Figure 6.7a shows that the data collected from the MRI, used as ground truth, lacks full constrictions for TTCD and TBCD, delivering

better performance only for LA. Tracking errors will invariably affect the system's performance when they occur often and systematically.

The second is related to the model design. The network is constructed to output at each time step t one curve $\hat{y}_{t,i} \in \mathbb{R}^{2\times 50}$ for each articulator i, and the learning objective is simply a reconstruction error concerning the true curve $y_{t,i}$, obtained from the tracking algorithm. The model will try to copy the ground truth, but some images will miss the contact as discussed in Chapter 5. The resulting model will output an average shape that finds a compromise between cases with and without the contact. Nothing in the learning objective enforces that the places of articulation are reached; we assume they will be learned from the data. However, given the network design, it is hard to introduce a minimization of the tract variables since the model can easily find a shortcut in the cost function by dedicating a few points in the prediction to satisfy that condition and use the remaining to minimize the reconstruction error, producing an unrealistic vocal tract shape.

4 Autoencoder-Based Vocal Tract Shape Synthesis from the Sequence of Phonemes to be Articulated

Section 3 discussed the challenges of modelling the vocal tract shape and how it impacts reaching some articulatory targets during speech. Estimating all of the points in the articulators' curves gives too much freedom to the model. The samples are highly correlated, and the model does not provide phonetically relevant parameters to control the vocal tract shape, making it impractical to analyze the impact that changes in the vocal tract have in the synthesized speech. Most importantly, enforcing critical articulator constraints is difficult. For instance, imposing contact between the tongue and the palate leads to artificial and inconsistent shapes. The model can easily find shortcuts to optimize the cost function without meaningfully improving the model.

In this Section, we divided the problem into two parts. Initially, an autoencoder was trained to learn a lower-dimensional latent representation of each vocal tract articulator. An ideal representation should be compact and composed of a few independent parameters meaningfully controlling the shapes of the articulators. This autoencoder can be interpreted as an articulatory model whose latent space represents a set of articulatory control parameters.

Then, a recurrent neural network was trained to estimate these lower dimensional articulatory parameters for each phoneme in the input sequence. The pre-trained autoencoder serves as an auxiliary network to train the phoneme-to-articulation model. In this phase, the autoen-

coder's weights were frozen and not updated; hence, the autoencoder does not learn.

Let $x \in \mathbb{R}^T$ be the input phonetic sequence and $y \in \mathbb{R}^{T \times N_{\texttt{art}} \times 2 \times N_{\texttt{samples}}}$ be the associated vocal tract shapes, where T is the temporal dimension, $N_{\texttt{art}}$ is the number of considered articulators, and $N_{\texttt{samples}}$ is the number of samples in the articulators curve. On the one hand, our previous approach was searching for the function f such that

$$\hat{y} = f(x)$$

which minimizes $d(y, \hat{y})$, where d is the point-wise Euclidean distance. On the other hand, the proposed approach searches for the functions g_{enc}, g_{dec} and h such that

$$z = g_{\texttt{enc}}(y)$$
$$\hat{z} = h(x)$$
$$\hat{y} = g_{\text{dec}}(\hat{z})$$

and $d_1(z, \hat{z})$ and $d_2(y, \hat{y})$ are minimal, where d_1 is the L_2 norm in the latent space and d_2 is the point-wise Euclidean distance in the output space. Here, g_{enc} and g_{dec} are the autoencoder's encoder and decoder networks, respectively, and h is a recurrent encoder-decoder network similar to the one from Section 3. Since the recurrent encoder-decoder network is limited to exploring the autoencoder's latent space, it cannot use shortcuts that spuriously minimize the loss function. Thus, we can impose constraints to the reconstructed shapes that enforce a phonetically relevant synthesis.

4.1 Methods

Autoencoder

Chapter 3 defined the autoencoder as a non-linear method for efficiently learning how to encode information. Compared to the PCA, the autoencoder has a higher representational power due to the non-linearities between its layers; nevertheless, it lacks essential characteristics such as orthogonality, statistical independence, and ranked components. Training a PCA-like autoencoder requires the minimization of the covariance in the latent space and a specific algorithm to rank the components by explained variance [116].

Our autoencoder mimics a few of these ideas while keeping the non-linear structure. Still, we cover only a few of these requirements. The encoder and the decoder presented in Figure 6.8 are formed by a sequence of linear layers with `ReLU` activation followed by an output linear

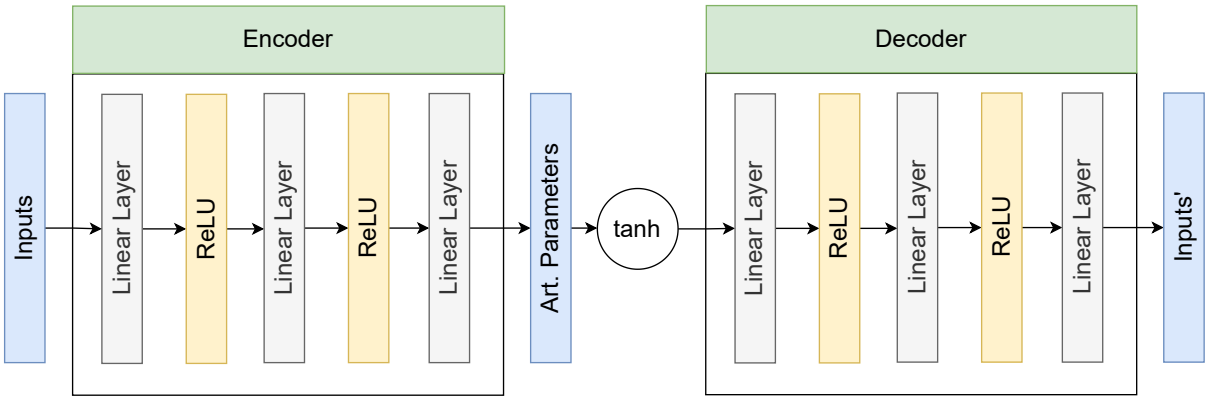


Figure 6.8: Autoencoder architecture.

layer. The hyperbolic tangent activation between the encoder and the decoder guarantees that the components are in the interval $[-1, 1]$, restricting the search space and improving model stability. We have one autoencoder model for each articulator, which does not share weights.

The autoencoder is trained using a loss function given by

$$
\begin{aligned}
z &= \texttt{tanh}(g_{\texttt{enc}}(x)) \\
\hat{x} &= g_{\texttt{dec}}(z) \\
\mathcal{L}_{\texttt{rec}}(x, \hat{x}) &= L_2(x, \hat{x}) \\
\mathcal{L}_{\texttt{cov}}(z) &= \texttt{cov}(z)^2 - \texttt{diag}(\texttt{cov}(z)^2) \\
\mathcal{L}(y, \hat{y}, z) &= \beta_1 \mathcal{L}_{\texttt{rec}}(y, \hat{y}) + \beta_2 \mathcal{L}_{\texttt{cov}}(z)
\end{aligned}
$$

where x and $\hat{x}$ represent the original and the reconstructed inputs, respectively, and z represents the autoencoder's latent space after the `tanh`, which we often refer to as articulatory parameters or autoencoder's components in this text. The autoencoder's latent space is represented by Z, the i^{th} component is denoted by z_i and $\texttt{dim}\ Z$ denotes the dimensionality of Z.

The primary metric for the autoencoder training is the reconstruction error, measured by the point-to-closest-point distance ($\texttt{P2CP}_{\texttt{mean}}$), presented in Chapter 4. A secondary evaluation method analyzes the impact of each component in the reconstructions. We vary each component from -1 to 1 while keeping all remaining components at value zero. Although this evaluation is subjective, it helps to determine whether the latent space meaningfully controls the vocal tract shape and can be used as an articulatory model.

The autoencoder was trained with the Adam optimizer [293] with the reduce learning rate on plateau scheduling policy. Even though we have one autoencoder for each articulator, and these networks do not share weights, they are trained jointly. Our hyperparameters choices are

Table 6.6: Hyperparameters of the autoencoder model training.

Hyperparameter	Value
$N_{\texttt{art}}$	10
$N_{\texttt{samples}}$	50
Batch size	12
Early-Stopping Patience	30
Weight Decay	10^{-5}
Learning Rate (LR)	10^{-4}
LR Sched. Patience	10
Sched. Reduction Factor	10
β_1	1.0
β_2	0.1
Total Number of Network Parameters	129 285

detailed in Table 6.6. In addition, we fit a PCA model from the `scikit-learn` library[4] to the same data and with the same number of components per articulator as our model as a baseline.

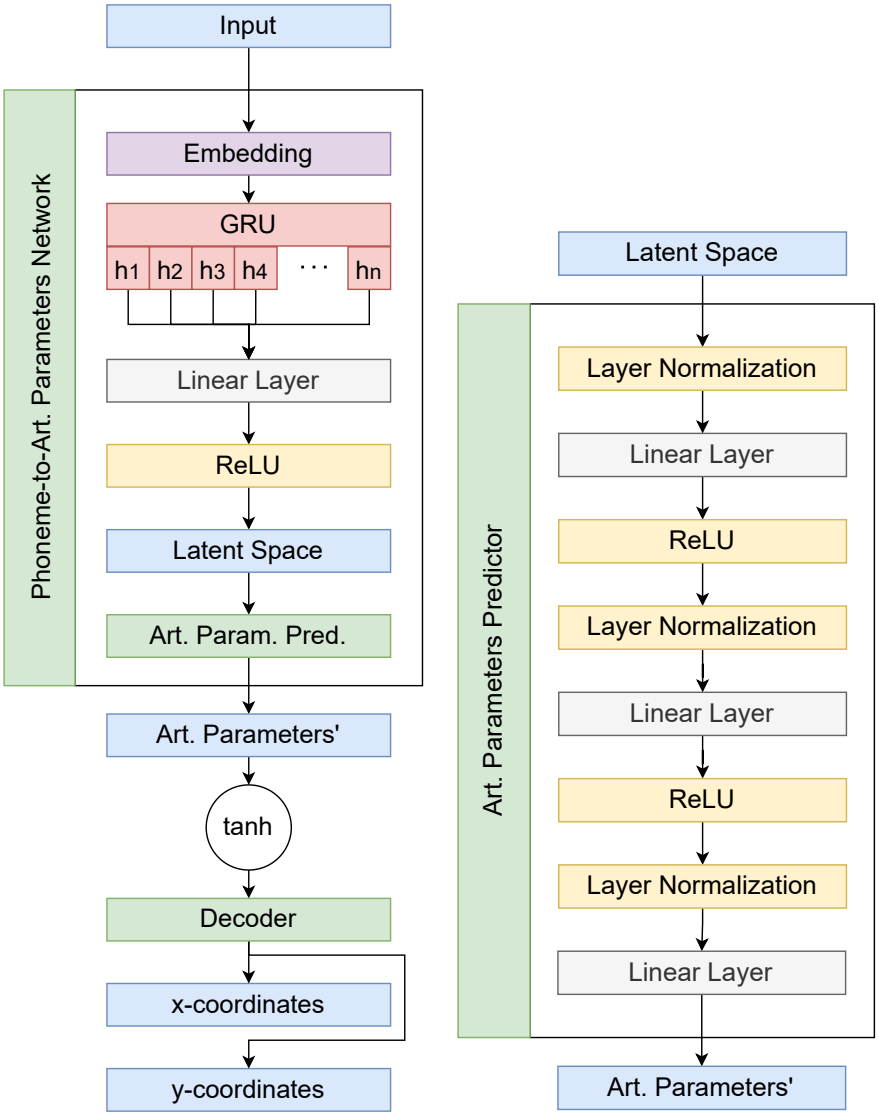
Autoencoder-Based Phoneme-to-Articulation

Figure 6.9 presents the architecture for the autoencoder-based phoneme-to-articulation network. Using the autoencoder as an intermediate articulatory model does not significantly change the architecture of the phoneme-to-articulation model from Section 3. The BiGRU model is still used, and the Articulatory Parameters Predictor is similar to the Articulator Predictor, except for the output layer. The main modification is that the phoneme-to-articulation model from Section 3 had one Articulator Predictor block per articulator, while the current model has only one Articulatory Parameters Predictor block that outputs the entire autoencoder's latent space.

The input to the network is the phonetic sequence with phoneme duration encoded as repetitions. The network outputs the vocal tract parameters for each time step. The predicted vocal tract shape is obtained by passing the predicted parameters to the decoder. The objective function is the most significant change compared to the model-free approach. The trivial objective function is the articulatory parameter prediction error, which we refer to as latent loss ($\mathcal{L}_{\texttt{latent}}$). Then, we can impose a cost on the vocal tract reconstruction error. We refer to this cost function as reconstruction loss ($\mathcal{L}_{\texttt{rec}}$).

However, the novelty is using the tract variables and phoneme-wise places of articulation to encourage the model to produce phonetically relevant vocal tract shapes. This cost is built by computing the minimal distance between articulator pairs for each time step. However, not all pairs are pertinent for all time steps. A binary mask is necessary to determine when each

[4] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.IncrementalPCA.html



(a) Autoencoder-Based Phoneme-to-Articulation

(b) Articulatory Parameters Predictor

Figure 6.9: Autoencoder-based phoneme-to-articulation architecture.

articulator pair is relevant for executing the phoneme in that time step. We refer to this loss function as critical loss ($\mathcal{L}_{\texttt{critical}}$).

The complete loss function for training the autoencoder-based phoneme-to-articulation is given by

$$
\begin{aligned}
z &= \texttt{tanh}(g_{\texttt{enc}}(y)) \\
\hat{z} &= \texttt{tanh}(h(x)) \\
\hat{y} &= g_{\texttt{dec}}(z) \\
\mathcal{L}_{\texttt{latent}}(z, \hat{z}) &= L_2(z, \hat{z}) \\
\mathcal{L}_{\texttt{rec}}(y, \hat{y}) &= L_2(y, \hat{y}) \\
\text{mindist}(u, v) &= \min_{\substack{i \in \{1,2,\ldots,N\} \\ j \in \{1,2,\ldots,N\}}} d(u_i, v_j) \\
\mathcal{L}_{\texttt{critical}}(\hat{y}) &= M \left[\text{mindist}_{\substack{t \in \{1,2,\ldots,T\} \\ i \in \{1,2,\ldots,N_{\text{art}}\} \\ j \in \{1,2,\ldots,N_{\text{art}}\}}} (\hat{y}_{t,i}, \hat{y}_{t,j}) \right] \\
\mathcal{L}(y, \hat{y}, z, \hat{z}) &= \beta_1 \mathcal{L}_{\texttt{latent}}(z, \hat{z}) + \beta_2 \mathcal{L}_{\texttt{rec}}(y, \hat{y}) + \beta_3 \mathcal{L}_{\texttt{critical}}(\hat{y})
\end{aligned}
$$

where y and $\hat{y}$ represent the target and the reconstructed vocal tract shapes, respectively, and z and $\hat{z}$ are the target and predicted articulatory parameters. M is a binary mask specifying which articulator pairs are critical for the execution of the phoneme x_t.

The autoencoder-based phoneme-to-articulation network was evaluated on the same basis as the model from Section 3 for a fair comparison, i.e., measuring the reconstruction error and tract variables. The models were trained with the Adam optimizer [293] with the reduce learning rate on plateau scheduling policy. Our hyperparameter choices are detailed in Table 6.7.

In summary, the main aspects of the proposed method are:

- **Inputs:** The input to the network is the same as in Section 3. A sequence of phonemes is represented by its index in a vocabulary, and the phoneme duration is encoded as token repetitions ($x \in \mathbb{R}^T$).

- **Autoencoder:** The autoencoder works as an articulatory model that learns articulatory parameters to reconstruct and control the vocal tract shape. Its architecture is detailed in Figure 6.8.

- **Articulatory Parameters:** It corresponds to the autoencoder's latent space, also re-

Table 6.7: Hyperparameters of the autoencoder-based phoneme-to-articulation training. The second part of table refers to the neural network hyperparameters.

Hyperparameter	Value
$N_{\texttt{art}}$	10
$N_{\texttt{samples}}$	50
Batch size	8
Early-Stopping Patience	30
Weight Decay	10^{-5}
Learning Rate (LR)	10^{-4}
LR Sched. Patience	10
Sched. Reduction Factor	10
β_1	0.5
β_2	3.0
β_3	1.0
Embedding Dim.	64
RNN Hidden Size	128
RNN Dropout	0.1
Predictor Hidden size	256
Total Number of Network Parameters	552 995

ferred to in this text as autoencoder's components. These parameters are equivalent to the PCA model's components.

- **Bidirectional GRU:** Similar to Section 3, the BiGRU allows the network to learn contextual dependencies in the input sequence.
- **Articulatory Parameters Predictor Block:** The architecture detailed in Figure 6.9 is used to predict the articulatory parameters from the latent space ($\hat{z} \in \mathbb{R}^{T \times \texttt{dim}Z}$).
- **Outputs:** The output of the complete system is the same as in Section 3. It is a sequence of articulator contours in the form $\hat{y} \in \mathbb{R}^{T \times N_{\texttt{art}} \times 2 \times N_{\texttt{samples}}}$ after the reconstruction by the decoder network.

4.2 Results

Autoencoder

The number of articulatory parameters and the reconstruction errors are presented in Table 6.8. The box plots from Figure 6.10 help to visualize, analyze, and compare these results more meaningfully. The PCA model achieved a reconstruction error below 1 mm for all articulators except the tongue. Conversely, the autoencoder obtained reconstruction errors below the PCA model for all articulators except the lower incisor and the soft palate. However, the differences were negligible in both cases.

Table 6.8: Number of components and reconstruction errors ($\texttt{P2CP}_{\texttt{mean}}$) for each articulator for the PCA and autoencoder. The best results are in bold.

Articulator	Number of Components	PCA $\texttt{P2CP}_{\texttt{mean}}$ (mm)	Autoencoder $\texttt{P2CP}_{\texttt{mean}}$ (mm)
Arytenoid Cartilage	4	0.88 ± 0.52	$\mathbf{0.50 \pm 0.19}$
Epiglottis Center Line	3	0.38 ± 0.18	$\mathbf{0.35 \pm 0.18}$
Lower Incisor	3	$\mathbf{0.01 \pm 0.00}$	0.02 ± 0.05
Lower Lip	4	0.49 ± 0.20	$\mathbf{0.46 \pm 0.20}$
Pharynx	2	0.62 ± 0.17	$\mathbf{0.56 \pm 0.17}$
Soft Palate Center Line	3	$\mathbf{0.40 \pm 0.14}$	0.43 ± 0.19
Thyroid Cartilage	2	0.51 ± 0.19	$\mathbf{0.45 \pm 0.19}$
Tongue	8	1.01 ± 0.22	$\mathbf{0.85 \pm 0.17}$
Upper Lip	4	0.48 ± 0.19	$\mathbf{0.38 \pm 0.14}$
Vocal Folds	2	0.83 ± 0.43	$\mathbf{0.54 \pm 0.25}$

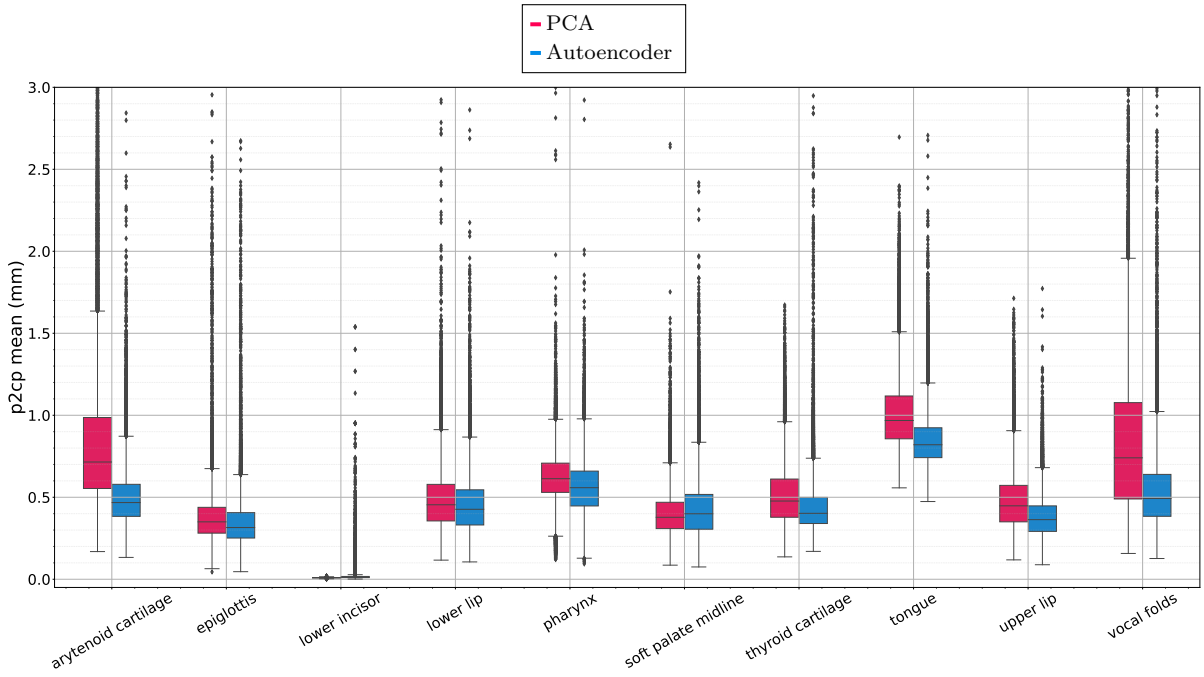


Figure 6.10: PCA and autoencoder reconstruction errors for each articulator in terms of the mean point-to-closest-point-error. The y-axis was clipped at 3 mm.

The nomograms presented in Figure 6.11 and Figure 6.12 show the effect of varying each autoencoder component individually in the $[-1, 1]$ interval. The gray lines correspond to articulators that are not affected by the component being analyzed.

Autoencoder-Based Phoneme-to-Articulation

The results presented here follow the same structure as those from Section 3. The main difference is that the model-free phoneme-to-articulation is the baseline for the current experiments.

Table 6.9 and Figure 6.13 presents the $\texttt{P2CP}_{\texttt{mean}}$ statistics for the model-free baseline and the autoencoder-based system for each articulator. The table presents the data in tabular form, while the figure shows a box plot to facilitate results visualization and comparison. Table 6.10 and Figure 6.14 presents Pearson's correlation statistics for each TV for the two approaches.

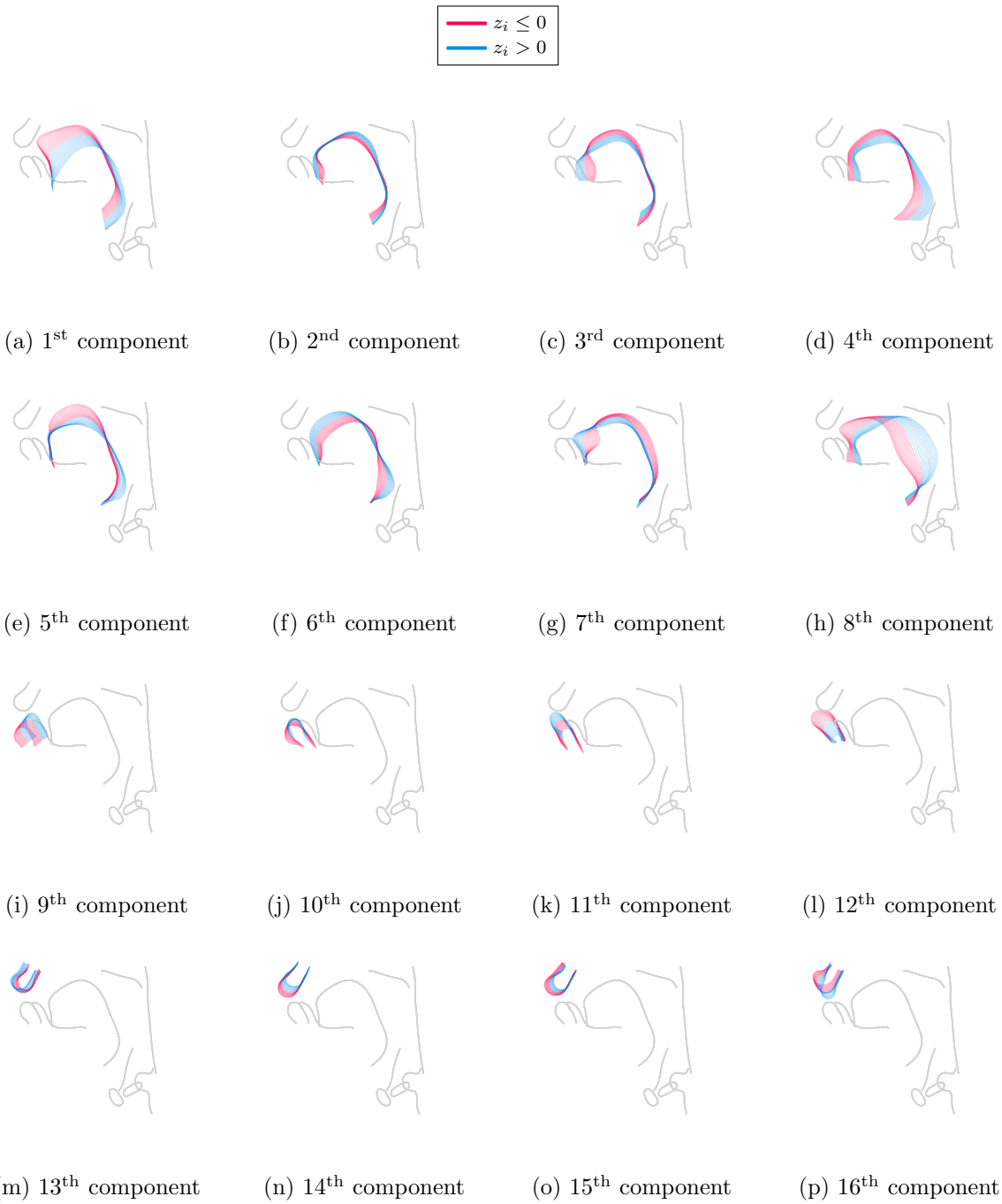
Figure 6.15 presents each TV's trajectories and corresponding tokens for one utterance in the test set using the autoencoder-based method. Additional samples are available in Section 2 of Appendix A. Figure 6.16 shows the ground truth, model-free prediction, and autoencoder-based prediction for the same utterance in the test set. As in the previous Section, the green arrows in Figure 6.6 indicate cases in which an articulatory target was reached, and the red arrows indicate the opposite.

4.3 Discussion

Autoencoder

Choosing the appropriate number of components to decompose each articulator is not trivial, and it gets progressively more complicated with the increased number of articulators and degrees of freedom of each articulator. Selecting a number that captures most of the variability without retaining spurious variance is necessary. In addition, the autoencoder's reconstruction error in Table 6.8 is directly related to the number of components, meaning that more components lead to a lower reconstruction error in the training set, and represents a performance lower bound for the vocal tract synthesis, i.e., even if the model makes perfect estimations of these control parameters, we shall still observe those reconstruction errors.

The errors from Table 6.8 show that the designed autoencoder presents outstanding reconstruction errors, lower than 1 mm on average for all articulators. The results are encouraging when compared to the baseline. The PCA is a robust and stable method for dimensionality reduction, providing a statistically independent parameter set that concentrates the most variability in the data. It is commonly used in articulatory speech research as an articulatory model. Even though the proposed autoencoder does not guarantee statistically independent



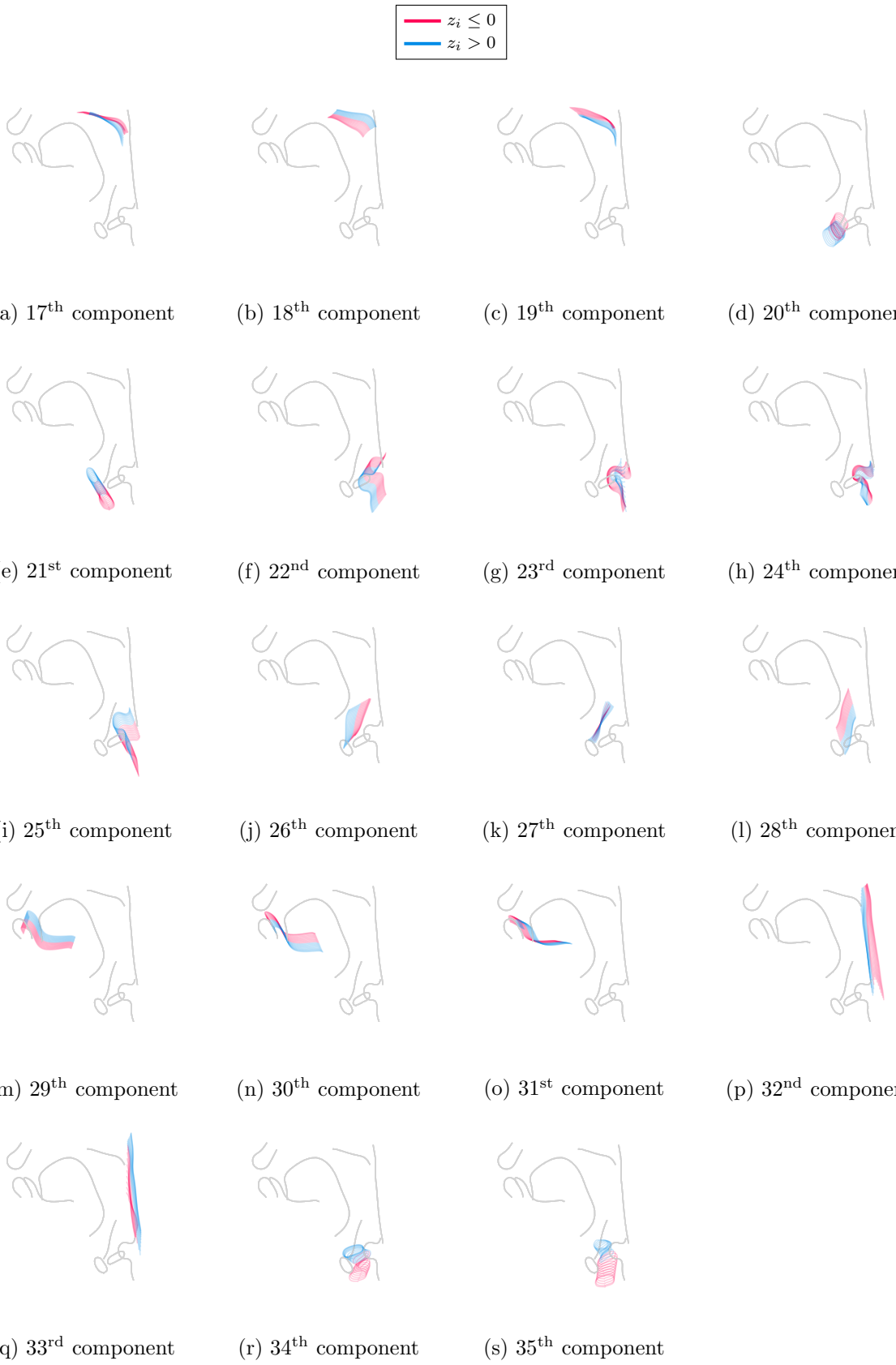
(a) 1st component (b) 2nd component (c) 3rd component (d) 4th component

(e) 5th component (f) 6th component (g) 7th component (h) 8th component

(i) 9th component (j) 10th component (k) 11th component (l) 12th component

(m) 13th component (n) 14th component (o) 15th component (p) 16th component

Figure 6.11: Nomogram displaying the effect of each component in the tongue, lower and upper lips's reconstruction.



(a) 17th component (b) 18th component (c) 19th component (d) 20th component

(e) 21st component (f) 22nd component (g) 23rd component (h) 24th component

(i) 25th component (j) 26th component (k) 27th component (l) 28th component

(m) 29th component (n) 30th component (o) 31st component (p) 32nd component

(q) 33rd component (r) 34th component (s) 35th component

Figure 6.12: Nomogram displaying the effect of each component in the reconstruction of the soft palate, thyroid cartilage, arytenoid cartilage, epiglottis, lower incisor, pharynx, and vocal folds.

Table 6.9: Reconstruction error ($\texttt{P2CP}_{\texttt{mean}}$) for the model-free and the autoencoder-based approaches. The best results are in bold.

Articulator	Method	$\texttt{P2CP}_{\texttt{mean}}$
Arytenoid Cartilage	Model-Free Phon.-to-Art.	1.77 ± 0.86
	Autoencoder-Based Phon.-to-Art.	$\mathbf{1.74 \pm 0.84}$
Epiglottis Center Line	Model-Free Phon.-to-Art.	1.55 ± 1.10
	Autoencoder-Based Phon.-to-Art.	$\mathbf{1.52 \pm 1.05}$
Lower Incisor	Model-Free Phon.-to-Art.	1.46 ± 0.78
	Autoencoder-Based Phon.-to-Art.	$\mathbf{1.45 \pm 0.78}$
Lower Lip	Model-Free Phon.-to-Art.	1.39 ± 0.67
	Autoencoder-Based Phon.-to-Art.	$\mathbf{1.37 \pm 0.67}$
Pharynx	Model-Free Phon.-to-Art.	$\mathbf{1.07 \pm 0.46}$
	Autoencoder-Based Phon.-to-Art.	1.11 ± 0.45
Soft Palate Center Line	Model-Free Phon.-to-Art.	$\mathbf{1.48 \pm 0.84}$
	Autoencoder-Based Phon.-to-Art.	1.62 ± 0.86
Thyroid Cartilage	Model-Free Phon.-to-Art.	$\mathbf{1.53 \pm 0.87}$
	Autoencoder-Based Phon.-to-Art.	1.54 ± 0.83
Tongue	Model-Free Phon.-to-Art.	$\mathbf{2.19 \pm 0.88}$
	Autoencoder-Based Phon.-to-Art.	2.34 ± 0.88
Upper Lip	Model-Free Phon.-to-Art.	$\mathbf{0.90 \pm 0.34}$
	Autoencoder-Based Phon.-to-Art.	0.94 ± 0.33
Vocal Folds	Model-Free Phon.-to-Art.	1.88 ± 1.02
	Autoencoder-Based Phon.-to-Art.	$\mathbf{1.71 \pm 1.98}$

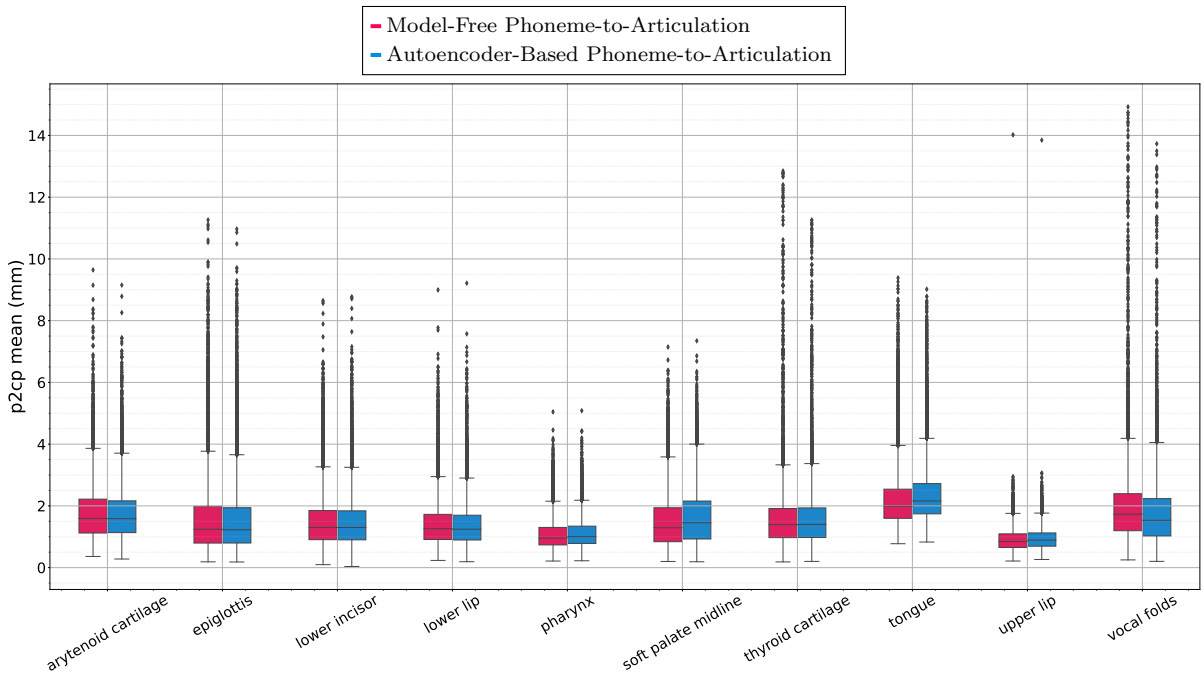


Figure 6.13: Box plots of the reconstruction error for the model-free and autoencoder-based approaches for each articulator.

Table 6.10: Correlations between the target and predicted tract variables trajectories.

Tract Valiable	Method	Correlation	Min Correlation	Max correlation
LA	Model-Free Phon.-to-Art.	0.80 ± 0.16	-0.41	0.95
	Autoencoder-Based Phon.-to-Art.	0.80 ± 0.14	-0.29	0.95
TTCD	Model-Free Phon.-to-Art.	0.86 ± 0.07	0.66	0.97
	Autoencoder-Based Phon.-to-Art.	$\mathbf{0.79 \pm 0.19}$	-1.00	0.95
TBCD	Model-Free Phon.-to-Art.	0.82 ± 0.08	0.44	0.94
	Autoencoder-Based Phon.-to-Art.	$\mathbf{0.80 \pm 0.08}$	0.44	0.96
VEL	Model-Free Phon.-to-Art.	0.74 ± 0.18	-1.00	0.90
	Autoencoder-Based Phon.-to-Art.	0.74 ± 0.08	0.43	0.91

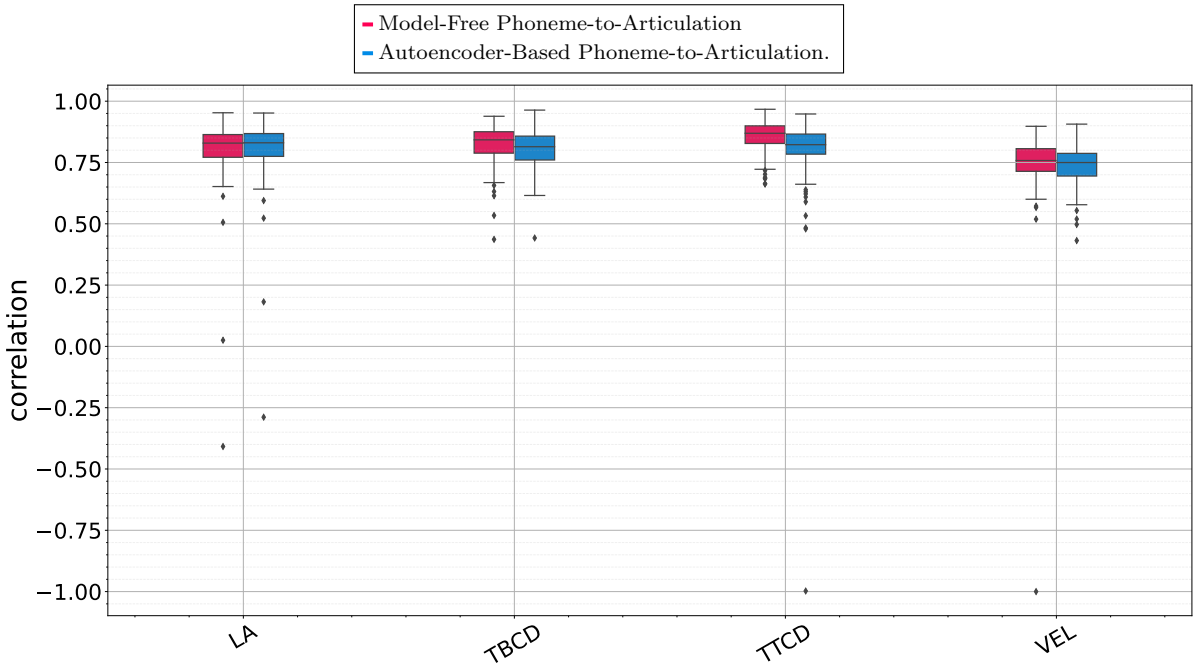
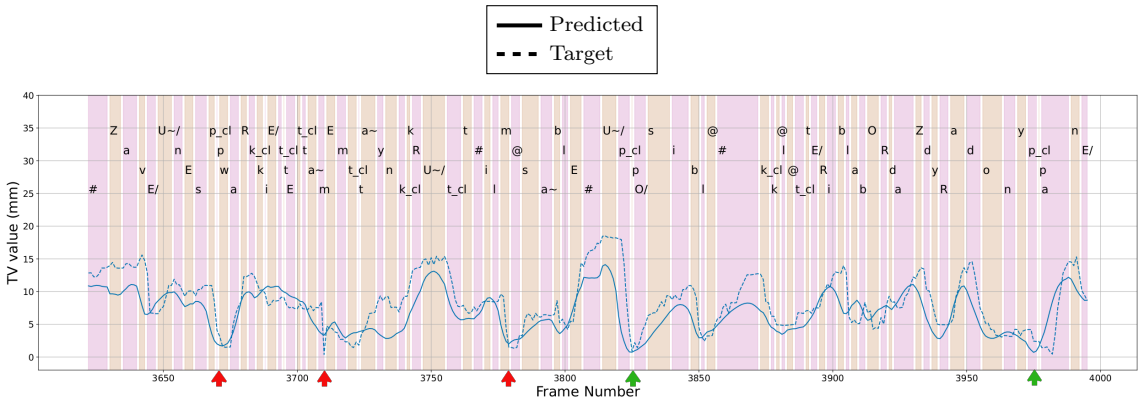
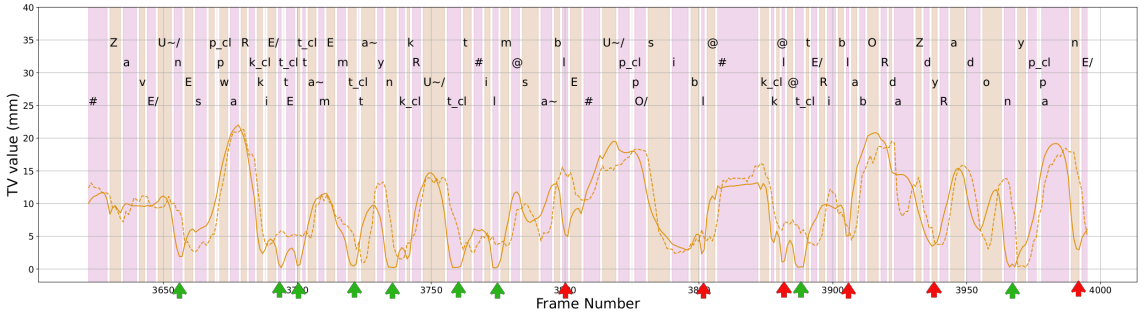


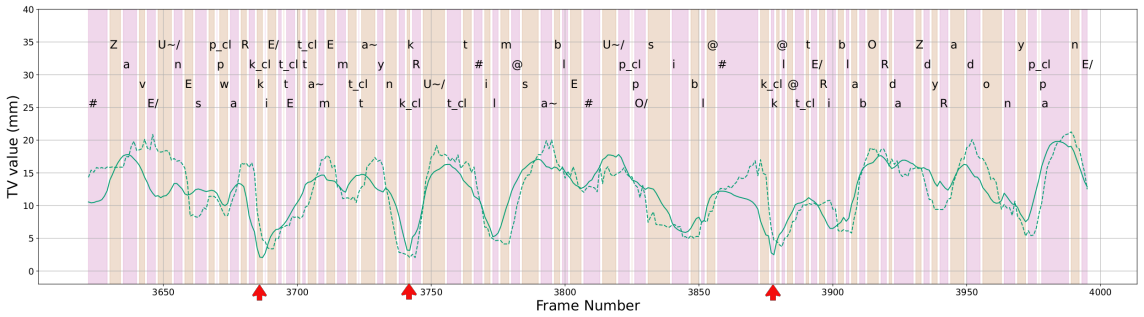
Figure 6.14: Box plots of the TVs correlations for the model-free and autoencoder-based approaches.



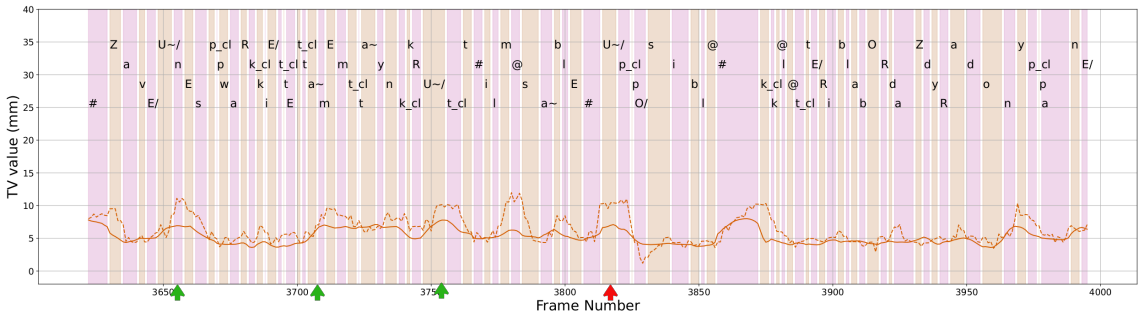
(a) LA



(b) TTCD

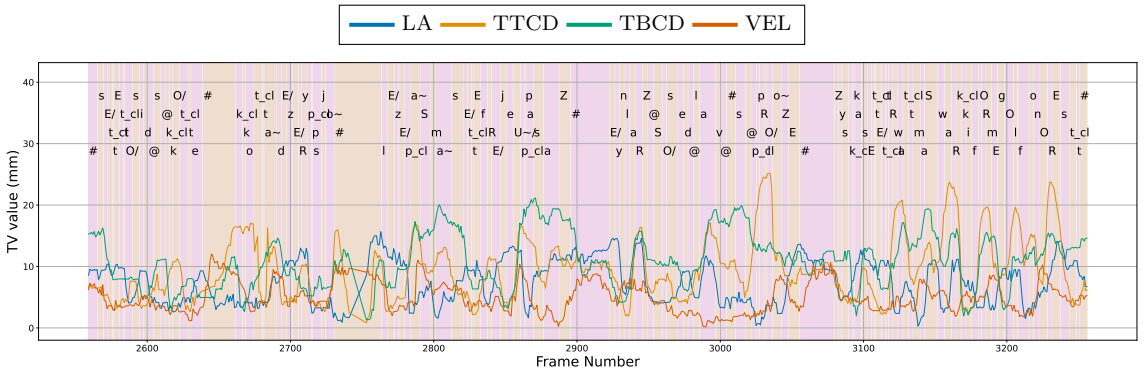


(c) TBCD

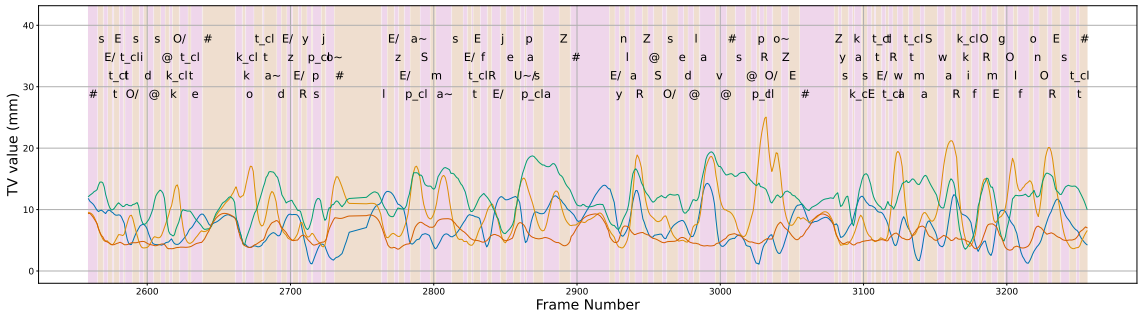


(d) VEL

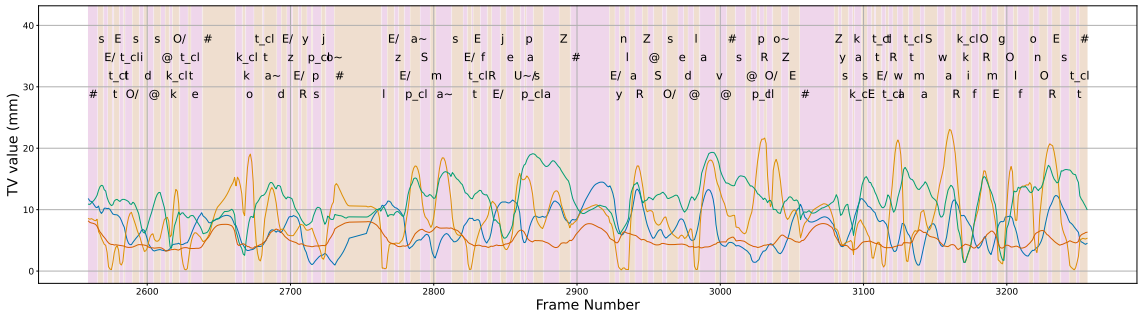
Figure 6.15: Autoencoder-based network and ground truth TV trajectories for the utterance *"J'avais un espoir qui était en même temps une crainte, il me semblait impossible que le terrible abordage du radeau n'eût pas anéanti."* Each image displays one tract variable. The corresponding phonemes are displayed in the top of the image. Green and red arrows indicate whether the model reaches an articulatory target or not, respectively. The alternating colors mark the onset and offset of each phoneme.



(a) Ground Truth



(b) Model-Free Phoneme-to-Articulation



(c) Autoencoder-Based Phoneme-to-Articulation

Figure 6.16: Ground truth, model-free prediction, and autoencoder-based prediction for the utterance *"C'était aussi de ce côté qu'au temps des éruptions, les épanchements s'étaient frayés un passage, et une large chaussée de laves se prolongeait jusqu'à cette étroite mâchoire qui formait golfe au nord-est."* The corresponding phonemes are displayed in the top of the image. The alternating colors mark the onset and offset of each phoneme.

and ranked components, it proved to be more valuable than the PCA as an articulatory model for the phoneme-to-articulation phase due to being bounded by the `tahn` function. A few experiments with a PCA-based phoneme-to-articulation system did not converge, primarily due to the unbounded PCA components.

In Ribeiro and Laprie [17], `dim` Z was chosen by testing different values and balancing the reconstruction capacity and the additional complexity. In that work, we used eight components for the tongue, which was kept in this work. However, we did not follow the same systematic approach for the remaining articulators due to the increased number of experiments that would be necessary without an equivalent benefit in the performances. We made a few explorations regarding the number of components and other hyperparameters but were limited to understanding the degrees of freedom of each articulator. Our experimentation showed that including extra components produced lower reconstruction errors at the expense of adding undesired artifacts to the nomograms from Figure 6.11 and Figure 6.12, which show how the compact set of articulatory parameters can meaningfully control the vocal tract shape, being a proper articulatory model for our purposes. With more parameters, the autoencoder fits spurious noise that does not contribute to our objectives.

Autoencoder-Based Phoneme-to-Articulation

The results for the autoencoder-based phoneme-to-articulation model are competitive when compared to the model-free approach from Section 3, especially considering that the autoencoder-based system contains a theoretical performance lower bound linked to the autoencoder's reconstruction error, which is not the case of our first method. However, we need to make a few considerations regarding the procedures.

The two systems presented very close results concerning the $\texttt{P2CP}_{\texttt{mean}}$ distances, differing on average in less than 0.5 mm for all articulators. Regarding TV trajectory correlations, both systems have very similar performance, with the autoencoder-based system outperforming the baseline for TTCD and TBCD. Then, evaluating which model is the best for the quantitative metrics is complicated.

The result is encouraging because it shows that a compact set of control parameters can still achieve valuable results. Additionally, since the indirect method is restricted to exploring the autoencoder's latent space, injecting critical articulator constraints into the reconstructed curves as loss functions is feasible. The most significant improvement is directly related to this characteristic. When comparing the same utterances for both models in Figure 6.6 and Figure 6.15, it is noticeable that the autoencoder-based approach yields more extreme TV values

for the associated phonemes, especially for TTCD.

Cases of under-articulation still occur, as shown by the red arrows. However, particularly for TTCD, the model tends to reach the articulatory target even when the information is not in the ground truth, as shown by the green arrows. When the ground truth misses a contact between two articulators, the autoencoder-based model deviates from the target trajectory to produce a more relevant vocal tract shape. In this sense, the critical loss injects prior knowledge from phonetics and linguistics into the model, helping to solve annotation errors or non-reached targets due to under-articulation.

The analysis of individual speech utterances synthesized by the models shows that they produce realistic vocal tract shapes concerning an external observer, with the model-free system resulting in more stable and temporally consistent utterances, as observed in the videos in the supplementary material. However, we did not propose a protocol to measure the mean opinion score because it requires independent evaluators familiar with the acoustics of the vocal tract. The struggle to rank the models raises concerns regarding the suggested evaluation metrics, questioning whether these metrics capture all the dimensions we want to measure in speech articulation synthesis.

Furthermore, it is essential to point out a few extra limitations of the autoencoder-based method:

1. The critical loss encourages correct places of articulation, but it does not guarantee them. There will be cases where the model fails to produce the necessary constrictions for proper acoustics.

2. The handcrafted critical loss function is complex to implement and use. Also, it only fits consonants since vocal tract closures do not characterize vowels. It would be preferred to have a method to learn the phonetic features implicitly from the data.

3. The autoencoder-based approach indirectly learns the vocal tract shape synthesis; hence, it is necessary to train an articulatory model (the autoencoder) first and then use it as the auxiliary network.

5 Evaluating Speech Articulation Synthesis Through Phoneme Recognition

In Section 3 and Section 4, we described how to synthesize the vocal tract shape conditioned in the sequence of phonemes to be articulated. However, assessing the quality of the synthesized

features remains a challenge. As discussed in Chapter 4, existing approaches include measuring the point-wise error between the predictions and the ground truth or the tract variables associated with each target phoneme. The first is easy to interpret, but the usage is limited due to inter- and intra-speaker variability. In contrast, the second fits the case of consonants well but does not suit vowels. Nevertheless, the two metrics were inconclusive regarding the proposed methods. The synthesized utterances from Section 3 seem more stable and consistent than those from Section 4, but the traditional metrics do not reflect this perception.

Inspired by recent works that perform phoneme recognition on MR images [219, 273] and EMA [274], we analyze speech articulations generated by a vocal tract shape synthesizer by their phonetic representations. We first train a phoneme recognizer on the acoustic signal as a baseline. Then, we train the recognizer on the real articulators' contours (true articulatory features). Since the mid-sagittal MRI does not include vocal fold excitation, we add a categorical encoding representing voicing information.

We quantify the information retained by the vocal tract contours by comparing the recognition error with the acoustic signal and the true articulatory features with and without voicing encoding. Next, the vocal tract shapes of the utterances in the test set were synthesized using the models from Section 3 and Section 4. These synthetic features with voicing encoding are input into the phoneme recognizer trained with the true articulations. The recognition error of this test exhibits how much phonetic information the synthesizer can reproduce.

If the true contours carry enough information, the representations learned from the acoustic and the articulatory data should be similar. The synthetic articulations should also exhibit a recognition performance comparable to the true articulatory features.

5.1 Methods

The acoustic features were obtained by computing the 80 log-Mel spectral features. The contours of the arytenoid cartilage, epiglottis center line, lower incisor, lower lip, pharynx, soft palate center line, thyroid cartilage, tongue, upper lip, and vocal folds presented in Figure 6.17 were concatenated to compose the articulatory features. The x- and y-coordinates form a 2-channel, 500-dimensional feature vector (10 articulators $\times$ 50 samples per curve). The synthetic articulatory features are obtained by inputting the test utterances into the synthesizers presented in Section 3 and Section 4, which return the synthetic articulatory features. The phonemes were grouped by their places of articulation for the evaluation, and the classes are described on Table 6.11.

The Deep Speech 2 [297] architecture inspires the phoneme recognizer. The network com-

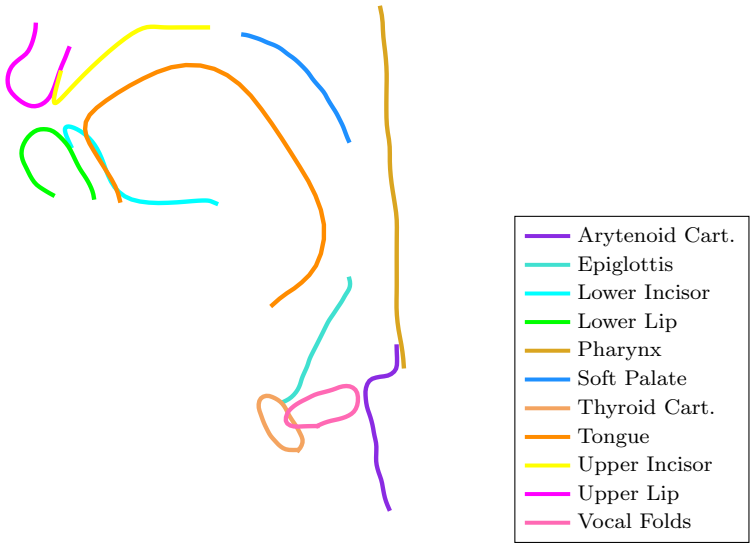


Figure 6.17: Articulatory features used for phoneme recognition plus the upper incisor, which is the reference for the coordinate system.

Table 6.11: Phonemes considered under each phonetic class. Phonemes with similar places of articulation are put grouped together.

Phonetic Classes	Phonemes
Dental	t, d, n, l, z, s
Labial	p, b, m, f, v
Palatal	k, g, ʒ, ʃ,
Front Vowels	i, e, ɛ, ɛ̃/œ̃, j
Back Vowels	u, o, ɔ, õ, w
Open Vowels	a, ã
Front Rounded Vowels	y, ø, œ, ɥ

Table 6.12: Hyperparameters of the phoneme recognition model training.

Hyperparameter	Value
$N_{\texttt{art}}$	10
$N_{\texttt{samples}}$	50
Batch size	12
Early-Stopping Patience	30
Weight Decay	10^{-5}
Learning Rate (LR)	10^{-4}
LR Sched. Patience	10
Sched. Reduction Factor	10
Gaussian Noise Std	10^{-4}
Total Number of Network Parameters (Acoustic Features)	297 362
Total Number of Network Parameters (Articulatory Features)	345 082

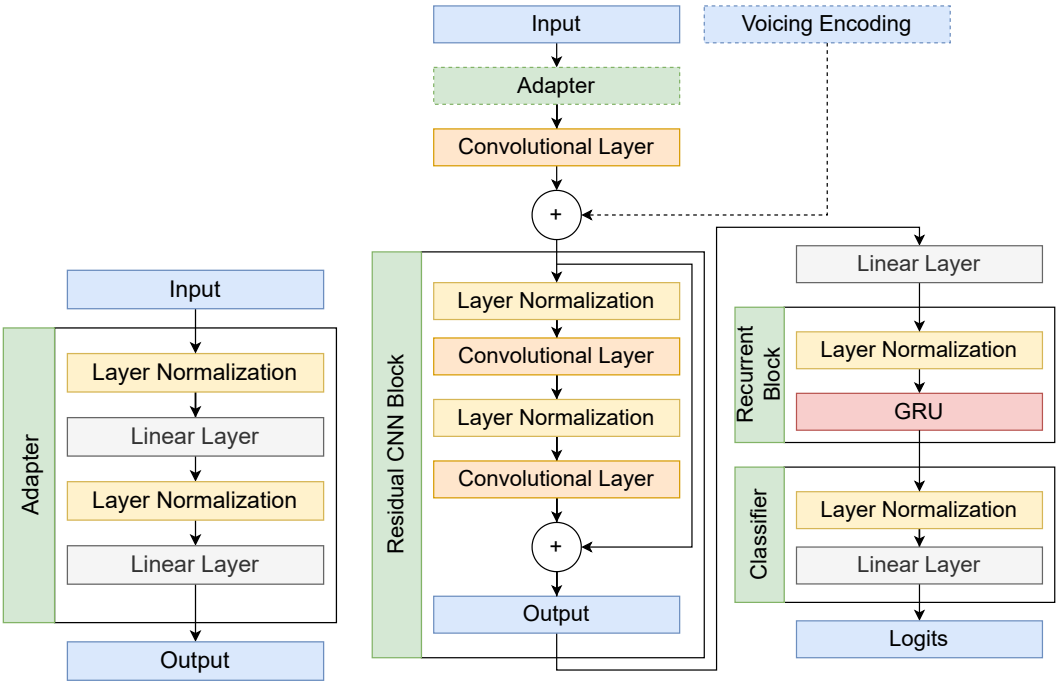
prises convolutional blocks with a residual additive connection between the inputs and the outputs, followed by recurrent blocks. Finally, a block of linear layers composes the classifier. To fit the articulatory features into the model, we prepend to the initial convolutional layer an adapter block formed by linear layers that convert the 500-dimensional tensor into an 80-dimensional feature vector. When voicing encoding was used, it was added to the outputs of the first convolutional layer. Figure 6.18 presents a schematic of the network architecture. Our implementation uses five residual convolutional blocks and three recurrent blocks.

The connectionist temporal classification (CTC) loss [298] was used as the learning objective, and the phoneme error rate (PER), measured in terms of the Levenshtein distance [299], is the evaluation metric. Furthermore, we computed the t-Distributed Stochastic Neighbor Embedding (t-SNE) [300] representations of the models' features calculated immediately before the classifier layer. The network was trained using the Adam optimizer [293] and the cyclic learning rate scheduler policy [108]. Additionally, we apply a slight Gaussian noise to the logits (model's outputs before the softmax) as a regularization strategy together with L_2 regularization. Our hyperparameter choices are detailed in Table 6.12.

In summary, the main aspects of the proposed method are:

- **Acoustic Features:** 80 log-Mel spectral features extracted directly from the audio recordings, forming a $T_{\texttt{acous}} \times 2 \times 80$ dimensional feature vector, where $T_{\texttt{acous}}$ is the length of the acoustic sequence. Ideally, these features would fully contain the phonetic information and be a recognition baseline.

- **True Articulatory Features:** Concatenation of ten vocal tract articulator contours directly extracted from the MRI films forming a $T_{\texttt{art}} \times 2 \times 500$ dimensional feature vector, where $T_{\texttt{art}}$ is the length of the articulatory sequence. These features measure the amount



(a) Adapter block

(b) Phoneme recognizer architecture with one residual CNN block and one recurrent block

Figure 6.18: Phoneme recognition network architecture. Activation and Dropout layers are omitted.

of phonetic information retained by the mid-sagittal contours.

- **Synthetic Articulatory Features:** Concatenation of the ten vocal tract articulators contours synthesized by the methods described in Section 3 and Section 4 forming a feature vector with the exact dimensions as the true articulatory features. These features measure the amount of phonetic information reproduced by the synthesizers.

- **Voicing Encoding:** Binary embedding informing if the target phoneme in time step t is voiced or unvoiced forming a $T_{\texttt{art}}$ dimensional feature vector. Since the mid-sagittal contours do not carry information regarding vocal fold vibration, the voicing encoding should help discriminate voiced and unvoiced phonemes.

- **Adapter Block:** A fully connected neural network that reduces the articulatory features dimensionality from 500 to 80 to fit the articulatory feature sets into the recognizer.

- **Phoneme Recognizer:** Neural network based on the Deep Speech 2 [297] architecture that transcribes the input features into the phonemes in the vocabulary. The model was trained using the CTC loss function [298].

- **Outputs:** The phonetic transcription predicted by the model.

Table 6.13: PER for the acoustic and articulatory features, with and without voicing encoding.

Feature Set	Voicing Encoding	PER
Acoustic Features	–	23.30
True Art. Features	No	23.65
Phon.-Wise Mean-Contour Art. Features	No	47.22
Model-Free Art. Features	No	24.34
Autoencoder-Based Art. Features	No	38.85
True Art. Features	Yes	21.66
Phon.-Wise Mean-Contour Art. Features	Yes	43.18
Model-Free Art. Features	Yes	**20.59**
Autoencoder-Based Art. Features	Yes	31.69

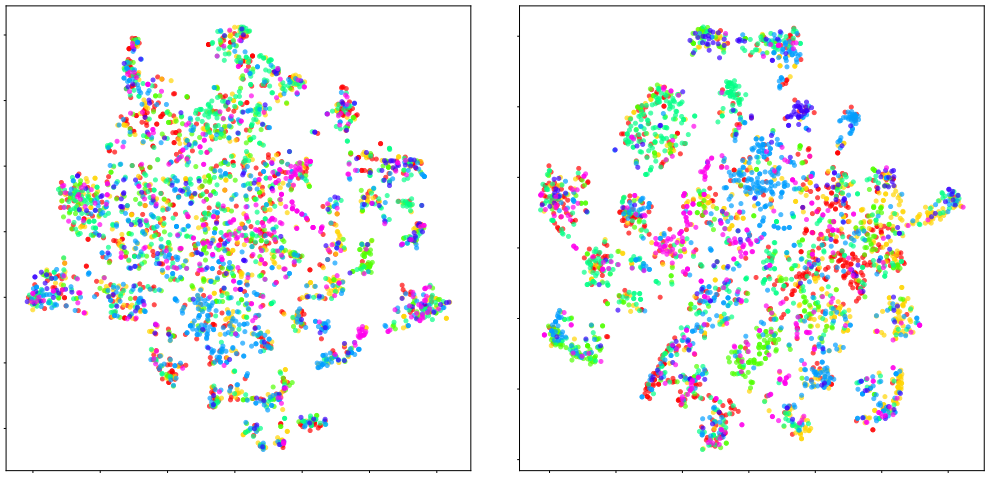
5.2 Results

Table 6.13 presents the PER for each feature set. Figure 6.19 shows the t-SNE plots of the phoneme representations learned by each model. The phonemes were grouped into their respective phonetic classes in Figure 6.19 to facilitate reading and visualization, and it includes only the phonemes listed in Table 6.11.

Figure 6.20 displays the ASR confusion matrix of the phoneme recognition, with phonemes grouped into their phonetic classes. Similarly to the confusion matrix used for traditional classification tasks, the rows represent the actual classes, and the columns represent the predicted classes. Each cell c_{ij} indicates the class i being substituted by class j; hence the main diagonal represents correct matches. The last column represents the deletions of each class, while the last row represents the insertions of each class. It is important to highlight that since the matrix is normalized by the true labels, the deletions column displays different information than the insertions row. While the element c_i in the deletions column shows the percentage of deleted tokens of class i, the element c_j in the insertions row presents the percentage of insertions corresponding to class j.

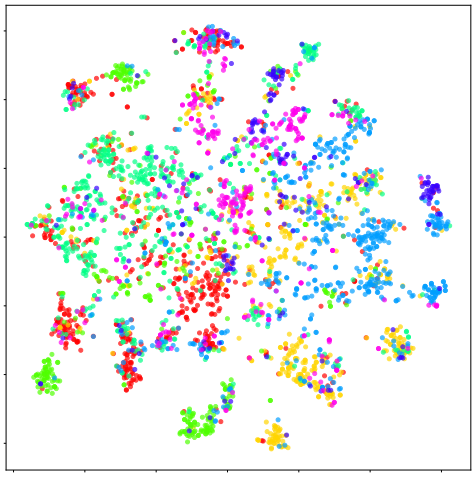
5.3 Discussion

The comparison between our models and the state of the art requires attention. The main benchmark for the task is the TIMIT dataset [301], and state-of-the-art models report a PER of 14.7 (wav2vec [302]) and 8.3 (wav2vec 2.0 [303]) on it. However, these models are much larger than ours and trained with massive data. Additionally, our recorded audio contains an intense MRI noise and is damaged by the denoising algorithm, contrarily to TIMIT, which has clean speech. Nevertheless, most importantly, outperforming these models is not our goal. Instead, we aim at quantifying the phonetic information retained by the articulatory features and the

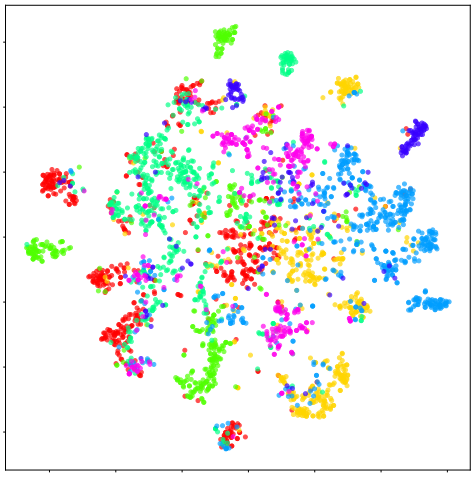


(a) Acoustic signal

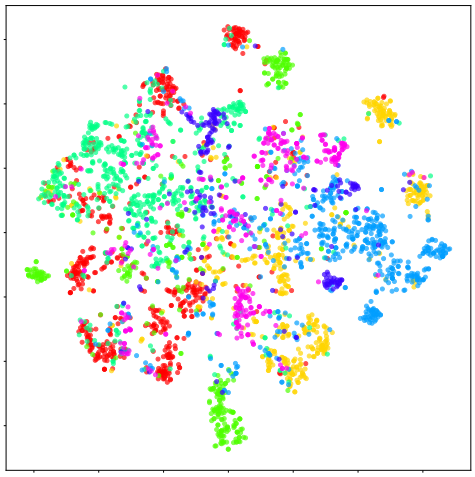
(b) True articulatory features



(c) True articulatory features + voicing



(d) Model-free articulatory features + voicing



(e) Autoencoder-based articulatory features + voicing

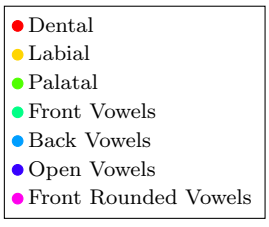
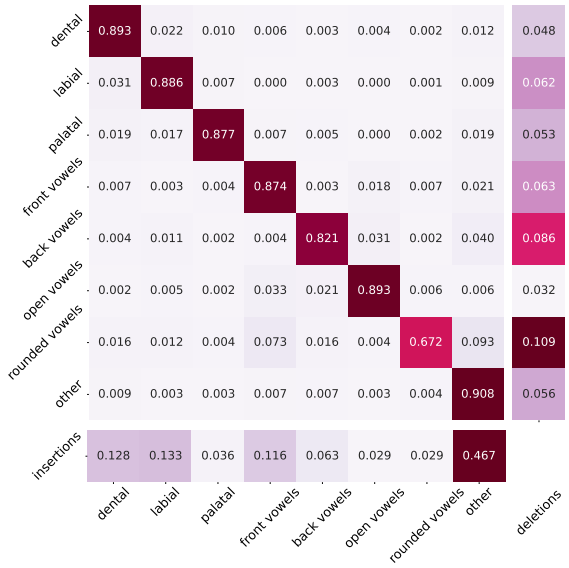
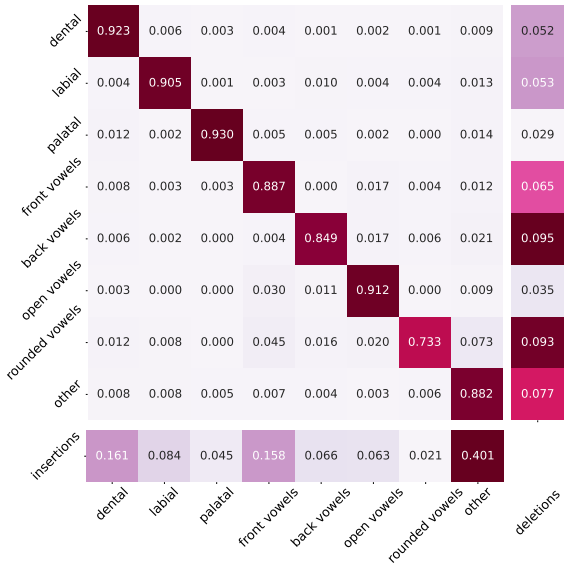


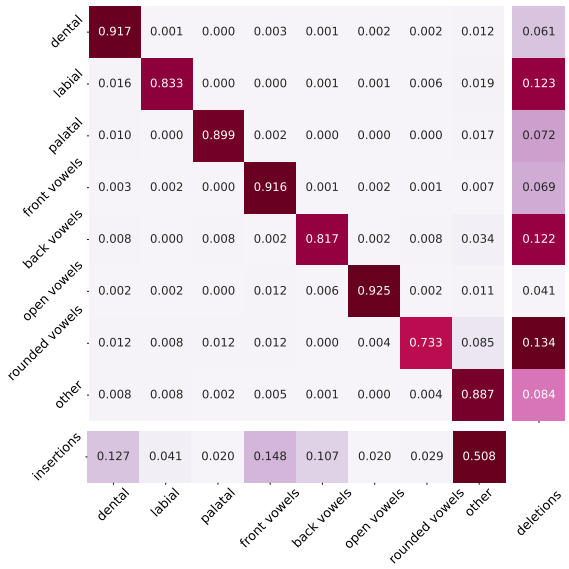
Figure 6.19: T-SNE plot of the phoneme representations for each feature set.



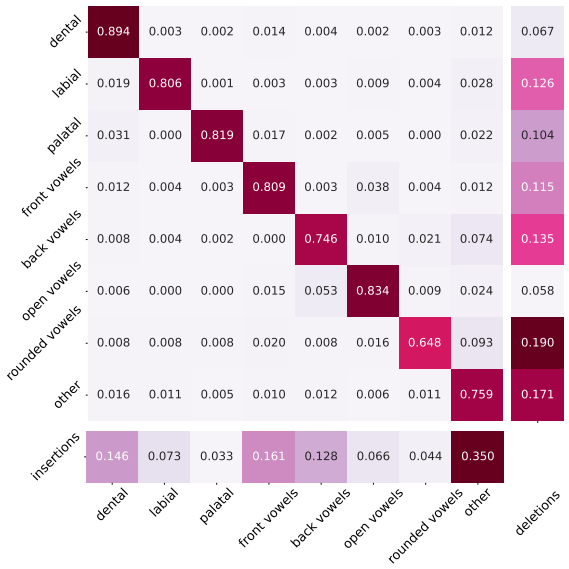
(a) Acoustic signal



(b) True articulatory features + voicing



(c) Model-free articulatory features + voicing



(d) Autoencoder-based articulatory features + voicing

Figure 6.20: Phoneme recognition confusion matrix normalized by the true labels. The true classes are shown in the rows while the predicted classes are displayed in the columns. The last column represents the deletions and the last row represents the insertions for each phonetic class.

one reproduced by the vocal tract synthesizer.

Moreover, we expect to use phoneme recognition as an evaluation metric for vocal tract synthesis. Still, the state-of-the-art models are a reference for judging if the recognizer's predictions are *good enough* to be used as a metric. That said, the models trained with the acoustic features and the articulatory features with voicing encoding resulted in a proper recognition compared to wav2vec but are still far from the results of wav2vec 2.0. Nevertheless, the results are satisfactory for our objective.

Table 6.13 shows that the recognition performance using the true articulatory features alone is indistinguishable from that of the acoustic features, which is a very satisfactory result since we expected that without the source information, the recognition would be much worse. Although surprising, the results are understandable. On the one hand, the articulators' contours extracted with the tracking method described in Chapter 5 are of high quality, showing outstanding performance in a multi-speaker setting. Despite the higher error in contact regions, the overall quality compensates for the errors. On the other hand, the substantial MRI noise in the acoustic features and the deterioration due to denoising contribute to a lower performance with the acoustic features.

Even if the articulatory features alone present performance very close to the acoustic signal, it is hard to believe that it retains the complete phonetic information. The vocal tract shapes lack source information, meaning unvoiced phonemes are indistinguishable from their voiced counterparts. After adding the voicing encoding to the feature set, the performance improved by 1.99 points. The comparison between Figure 6.19b and Figure 6.19c corroborates this idea since the groups formed by the articulatory features with voicing encoding are more evident than those without the source information.

Unsurprisingly, the phoneme-wise mean contour presents inferior recognition performance, which is expected due to the model's simplicity, which does not account for contextual information. The PER using the synthetic vocal tract shapes from the model-free approach with source information is outstanding. The recognition performance has a lower PER than all other feature sets, including the true articulatory features with voicing encoding, even if the latter is the same features used during training. Even if the model-free articulatory features are of high quality and the vocal tract shapes are realistic, the result is surprising. The reason might be that the articulatory synthesizer filters out noise in the true features, generating cleaner speech articulations. Conversely, the recognition performance using the autoencoder-based articulatory features is lower, only beating the mean-contour features. Even if the model-free and autoencoder-based systems presented very competitive results so far, we see that the PER can

discriminate the two models more meaningfully. The phoneme recognition captures the initial impression that the model-free system yields more consistent speech articulations.

Figure 6.19 shows that, with the addition of voicing encoding, the articulatory features form apparent groups in the embedding space that are not seen even with the acoustic features even if the recognition is not included in the synthesizers' optimization procedure. The PER and the feature embeddings corroborate the quality of the synthesized vocal tract shapes.

We need to address the issue of reaching the correct places of articulation. Section 3 discussed the difficulty of achieving proper dental, palatal, and labial constrictions. It should not be a surprise that the model has an exceptionally high deletion rate for labials and palatals (Figure 6.20c), which is not observed with the true articulatory features (Figure 6.20b) and the acoustic features (Figure 6.20a). In addition, we observe a high deletion rate for front rounded vowels, which is understandable since the mid-sagittal vocal tract shapes lack lip rounding.

The confusion matrix for the autoencoder-based system (Figure 6.20d) retains high deletion rates for the dental, labial, and palatal phonemes even though we observed an improvement in these places of articulation. Since the deletion rates with the autoencoder-based articulatory features are higher for all phonetic classes, it is unclear whether the high deletion rates for these specific classes are due to poor recognition performance or lack of proper articulatory constrictions.

6 Conclusion

This Chapter explored speech articulation synthesis conditioned on a phonetic sequence in French. We proposed three methods. The first is a straightforward baseline system to permit an initial comparison. Then, we proposed a model that directly maps the phonemes to the articulations without any intermediate articulatory model. This method resulted in outstanding vocal tract shapes, with a PER lower than the articulatory features extracted directly from the MRI, showing that it is the one that reproduces the most phonetic information.

However, direct mapping prohibits injecting prior phonetic knowledge into the model through its loss function. Additionally, the absence of an explicit articulatory model can be prohibitive for many research areas, such as speech motor control. Thus, we proposed a novel approach that trains the phoneme-to-articulation with the intermediation of an articulatory model.

First, we designed PCA-based and autoencoder-based articulatory models to encode vocal tract shapes extracted from the RT-MRI films into meaningful articulatory parameters. The autoencoder presented a superior reconstruction capacity than the PCA. Then, we trained phoneme-to-articulation mapping to predict the articulatory parameters. The final articulator

contours are obtained by inputting the predicted parameters into the articulatory model's decoding path. This approach shows the capacity to produce shapes with low errors. However, the synthesized utterances were less stable and temporally consistent than the model-free system (see supplementary videos), reducing the phoneme recognition performance.

Along with our research, we found evaluating and comparing our models difficult. The metrics available so far were objective but only captured some of the dimensions we desired. Point-wise metrics penalize models that learn various possible articulations and do not fit intra-speaker variability and, consequently, the multi-speaker cases. The tract variables alternatively measure the dynamics of speech and the interaction between articulators but might not fit vowels well. Thus, we dedicated the last part of the work to developing an evaluation system that would encourage the model to synthesize an intelligible articulatory set regardless of the speaker and that can fit a more sizeable phonetic context.

We trained a phoneme recognizer to transcribe articulatory features into phonemes. Our experiments showed that the true articulatory features extracted from the RT-MRI are phonetically rich and carry even more information than the acoustic baseline. This observation is compatible with our laboratory experience. The recorded audios are deficient due to the enormous MRI noise and the denoising algorithm. For instance, the vocal tract contours were more recognizable than some audio fragments for a trained phonetics and vocal tract acoustics professional.

Nevertheless, it is unknown to us which features the phoneme recognizer learns. In the future, it is desirable to perform further exploration of the recognizer. Explainability is a research topic that has concentrated much attention in recent years, and it would be interesting to use these frameworks to understand if the essential features for the recognition align with the phonetics literature.

Furthermore, we have focused our research on a single speaker. Extending it to the multi-speaker setting is desired. Since phoneme recognition is a speaker-independent function by design, a fruitful research line is to investigate how it could be introduced in the training pipeline to perform implicit speaker normalization.

Chapter 7

Conclusion

I did it for me.

Walter White

Vince Gilligan

1 Summary

Articulatory synthesis of speech is a multifaced problem. The literature typically covers the direct problem, articulatory speech synthesis, or its inverse form, acoustic-to-articulatory inversion. In the first case, articulatory data is the input to reconstruct the speech signal. Alternatively, articulatory copy synthesis tries to optimize control parameters such that the synthesized signal mimics the reference. In the second case, the acoustic signal is the input to recover speech articulations.

In this thesis, we proposed a third facet of articulatory synthesis. Our task, phoneme-to-articulation, uses the phonetic sequence to predict the vocal tract shape. This mapping has many applications in speech production, such as speech therapy and L2 learning. In addition, the capacity to recreate speech articulations taking coarticulation into account can potentially improve speech synthesis by creating a link with the speech production process, allowing a more natural signal.

The challenges surrounding phoneme-to-articulation include the difficulty of collecting and processing relevant data, which requires fine annotations both at the articulatory and phonetic levels, guaranteeing the physical realism of the generated vocal tract shapes, injecting relevant phonetic information into the model, and dealing with multiple sources of variance in speech. We designed innovative solutions to handle these problems and opened many fruitful directions for future work, which will be discussed next.

2 Main Contributions

Chapter 5 presented our approach to the design of a reliable and robust system to segment and track vocal tract articulators in real-time MRI. Nevertheless, medical imaging is challenging, particularly (RT-)MRI. Low spatial resolution means that even a single pixel error can be significant, and the low frame rate compared to speech rate introduces uncertainty about the actual position of fast-moving articulators. Additionally, different artifacts may be present in the MRI depending on the method used to sample the Fourier space, which can reduce the generalization capacity of the system.

The most significant obstacle in our research was obtaining accurate image annotations. It is difficult to estimate how many human hours were spent annotating and correcting semi-automatic annotations for training the system. Data annotation significantly impacts the performance of machine learning systems. Our results achieved the expected quality only because we dedicated sufficient effort to this step. We hope that releasing these annotations and the MRI movies will benefit other research teams in developing even more robust systems.

Despite these challenges, we successfully designed a method to extract vocal tract articulator contours in RT-MRI. Our system achieved outstanding performance, being stable, fast to train, and with a reasonable inference time, which allowed us to process a dataset of 548 000 images with minimal human interaction. Part of the dataset (`ASD1`) is already available, while we expect to release the other part (`ASD2`) soon. Furthermore, we have made the code for reproducing our experiments publicly available and distributed the system as an installable `Python` package. Our objective is to make it a publicly-available tool to boost the research in articulatory models, enabling the processing of large MRI datasets. Hopefully, this software will continue to evolve, with the perfection of the current features and the development of new ones.

Section 3 and Section 4 from Chapter 4 detailed our approaches to the main goal of the thesis. The first step was to develop a baseline system to compare our work. The literature in this area is broad, but it is rare to find approaches that follow our established protocols regarding data sources, annotations, and phonetic alignment. Compared to many other machine learning tasks, there are few vocal tract shape synthesis benchmarks, and well-annotated and curated datasets are scarce. A straightforward way to achieve a proper baseline was to use the average contour per phone interval.

Then, two methods were proposed. The first method directly predicts the vocal tract contours conditioned by the input phonetic sequence. This approach was the most successful in our final evaluation, generating the most realistic shapes. However, it also had some drawbacks. The predicted vocal tract shapes often miss important articulatory cues such as lip closure and

contact between the tongue and the alveolar region or the hard palate. Additionally, the system is difficult to control due to the absence of meaningful control parameters, which limits its impact in some research areas.

The second method models phoneme-to-articulation mapping with an articulatory model. The autoencoder showed an improved reconstruction capacity compared to the PCA benchmark. However, most importantly, it showed an improved power to control the vocal tract shape with a very compact set of articulatory parameters. The autoencoder-based phoneme-to-articulation mapping achieved comparable performance to the model-free approach in the traditional metrics. It also showed a better fit to critical articulation constraints. However, we observed that the synthesized shapes are less stable and temporally consistent, which may be the reason for the higher PER in the following experiments.

Finally, Section 5 from Chapter 4 detailed our approach to developing a metric for evaluating synthesized speech articulations. We were concerned that the traditional metrics, such as the point-to-closest point distance and correlation, were insufficient to evaluate our models.

We focused on using phoneme recognition to learn a function to measure synthetic shapes to address these challenges. Our experiments showed that the articulatory features extracted from RT-MRI carry enough phonetic information, showing a PER lower than the acoustic features. This result aligns with our understanding that the speech recordings are noisy. Moreover, our experiments showed that the phoneme recognizer has a lower error rate when testing the synthesized features then with the articulatory features extracted from the MRI, even if it was trained with the true articulatory features. This result was surprising but understandable, as the synthesized articulations are stable, filtering out noise from the automatic annotations.

By the end of this thesis, we are confident that we successfully achieved our three primary objectives. We have made significant contributions to the field of speech articulation research. We expect that the proposed methods will allow researchers to study the dynamics of the vocal tract in finer detail, which may lead to a better understanding of speech production. We are excited to see how other researchers will use our work in the future.

3 Directions for Future Work

Segmentation of Vocal Tract Articulations

The articulatory tracking system must include the upper and lower incisors to complete the vocal tract shape. Currently, we use image correlation to annotate this region, but this method requires human interaction and the adjustment of several hyperparameters. It would be desir-

able to include these rigid articulators in the pipeline to enhance the impact of our articulatory tracking system.

Furthermore, it is necessary to improve the management of contact regions. We have experimented with many unsuccessful approaches. One promising strategy is to give a higher weight to pixels close to the alveolar area. The challenge is to distinguish between cases where the contact happened (/t, d, n, l/) and those where there is a constriction without a contact (/i, y/). However, this would require processing the upper incisor before the tongue, reducing the parallelization of tasks and the system's efficiency. Another promising direction would be to train a neural network-based post-processing function that takes as input the segmentation mask and outputs the articulator contour. Other applications have used a similar approach, such as computing the vocal tract center line [23].

Speech Articulation Synthesis

The developed system for speech articulation synthesis does not have physical constraints, such as preventing articulators from overlapping in space. We expected these constraints to be learned directly from the data, which is partially true. The occurrence of overlaps in the synthetic utterances is negligible and does not significantly disturb the final vocal tract air column. However, introducing these constraints into the model would significantly increase its reliability.

Another area of future research is improving the synthesized articulations' physical realism. The autoencoder model used in this study encourages constrictions at appropriate places of articulation but does not guarantee them. Therefore, introducing these guarantees in the system would be a fruitful research topic.

Even though the autoencoder-based phoneme-to-articulation underperformed compared to the model-free method, it provides advantages that make it the preferred method for the continuation of this research. Compared to the previous methods, the access to articulatory control parameters and possibility to control the reconstructed curve without incurring unnatural shapes are relevant advantages. Therefore, it would be interesting to investigate how to improve this model, producing more stable and natural articulations.

Along with this thesis, we focused on recurrent neural networks. A few experiments were executed with the LSTM and the GRU, and we decided to use the latter in the final models. Recurrent networks fit our problem well because they are easy to train and require less data. Coarticulation modeling does not require learning long-term dependencies beyond the capacity of RNNs. A phoneme has an articulatory impact on its neighbors, but not in phonemes distant

in the sequence. Nevertheless, it would be interesting to explore transformer-based models on coarticulation modeling.

Observing how the transformer model aligns the input phonetic sequence with the target articulations and investigating the attention map at the attention head level could be fruitful. Typically, each attention head receives a fraction of the input sequence. This design is advantageous because it is possible to investigate how each region of the articulator aligns with each phoneme. For instance, the tongue tip is expected to have a higher attention weight with dental phonemes, and the tongue dorsum has a more substantial alignment with palatal phonemes.

Phoneme Recognition for Speech Articulation Synthesis

There are many exciting possibilities in phoneme recognition research with articulatory features. One of the most relevant possibilities is to understand the predictions made by the phoneme recognizer. The phonetics and linguistics literature define the articulatory targets for the utterance of each phoneme. Therefore, it would be essential to understand and ensure that the phoneme recognizer uses these features in the emission of each phoneme, allowing the use of the phoneme recognizer to evaluate if the synthesizer can reproduce these articulatory targets.

Another possibility is to use phoneme recognition in the training feedback loop of phoneme-to-articulation. The phoneme recognizer is a speaker-independent function by design; thus, data availability is the only obstacle to the multi-speaker setting. Using phoneme recognition in the training feedback loop could build a speaker-independent cost function that performs implicit speaker normalization, enabling the training of phoneme-to-articulation models on a larger dataset of speakers, which would improve the generalization performance of the models.

The question is whether phoneme recognition is enough to train phoneme-to-articulation mapping or if point-wise metrics are still required. Nevertheless, even if the traditional loss functions are still necessary, a recognition loss may still be helpful by injecting relevant phonetic information into the model without needing handcrafted loss functions, helping to synthesize more realistic vocal tract shapes.

Bibliography

[1] Yuval Noah Harari. *Sapiens: A brief history of humankind*. Random House, 2014.

[2] Sethu Karthikeyan, David A Puts, Toe Aung, Jennifer K Link, Kevin Rosenfield, Alexander Mackiel, Allisen Casey, Kaelyn Marks, Michele Cristo, Jenny Patel, et al. Articulatory effects on perceptions of men's status and attractiveness. *Scientific reports*, 13(1):2647, 2023.

[3] Christian Gottlieb Kratzenstein. *Tentamen resolvendi problema ab Acad. Petropolit. 1780 propositu qualis sit natura litterarum vocalium a, e, i, o, u.* 1781.

[4] Christian Korpiun. Kratzenstein's vowel resonators-reflections on a revival. In *HSCR@ INTERSPEECH*, pages 52–59, 2015.

[5] Google Arts & Culture. The 'Kempelen" speaking machine. URL https://artsandculture.google.com/story/2QUB7hLe64FKJA.

[6] Rubeena A Khan and Janardan Shrawan Chitode. Concatenative speech synthesis: A review. *International Journal of Computer Applications*, 136(3):1–6, 2016.

[7] Christine H Shadle and Robert I Damper. Prospects for articulatory synthesis: A position paper. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[8] Sangramsing Kayte, Monica Mundada, and Jayesh Gujrathi. Hidden markov model based speech synthesis: A review. *International Journal of Computer Applications*, 130(3):35–39, 2015.

[9] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077, 2020.

[10] Peter B Denes and Elliott N Pinson. The speech chain (baltimore, 1963). In *DB Fry," Automatic Speech Recognition," in Proceedings of the Fourth International Congress of Phonetic Sciences (The Hague, 1962)*, volume 314, pages 168–170, 1962.

[11] Shrikanth Narayanan. Bridging Speech Science and Technology — Now and Into the Future. In *Proc. INTERSPEECH 2023*, page 1, 2023.

[12] Christina Hagedorn, Michael Proctor, Louis Goldstein, Stephen M Wilson, Bruce Miller, Maria Luisa Gorno-Tempini, and Shrikanth S Narayanan. Characterizing articulation in apraxic speech using real-time magnetic resonance imaging. *Journal of Speech, Language, and Hearing Research*, 60(4):877–891, 2017.

[13] Christina Hagedorn, Jangwon Kim, Uttam Sinha, Louis Goldstein, and Shrikanth S Narayanan. Complexity of vocal tract shaping in glossectomy patients and typical speakers: A principal component analysis. *The Journal of the Acoustical Society of America*, 149(6):4437–4449, 2021.

[14] Christina Hagedorn, Yijing Lu, Asterios Toutios, Uttam Sinha, Louis Goldstein, and Shrikanth Narayanan. Variation in compensatory strategies as a function of target constriction degree in post-glossectomy speech. *JASA Express Letters*, 2(4), 2022.

[15] Ming Li, Jangwon Kim, Adam Lammert, Prasanta Kumar Ghosh, Vikram Ramanarayanan, and Shrikanth Narayanan. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. *Computer speech & language*, 36:196–211, 2016.

[16] Gokul Srinivasan, Aravind Illa, and Prasanta Kumar Ghosh. A study on robustness of articulatory features for automatic speech recognition of neutral and whispered speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5936–5940. IEEE, 2019.

[17] Vinicius Ribeiro and Yves Laprie. Autoencoder-Based Tongue Shape Estimation During Continuous Speech. In *Proc. INTERSPEECH 2022*, pages 86–90, 2022.

[18] V. Ribeiro, K. Isaieva, J. Leclere, PA Vuissoz, and Y. Laprie. Towards the Prediction of the Vocal Tract Shape from the Sequence of Phonemes to be Articulated. In *Proc. INTERSPEECH 2021*, pages 3325–3329, 2021.

[19] Vinícius Ribeiro, Karyna Isaieva, Justine Leclere, , Jacques Felblinger, Pierre-André Vuissoz, and Yves Laprie. Automatic segmentation of vocal tract articulators in real-time magnetic resonance imaging. *Available at SSRN 4192628*.

[20] Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Pierre-André Vuissoz, and Yves Laprie. Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated. *Speech Communication*, 2022. ISSN 0167-6393.

[21] Y. Laprie, V. Ribeiro, K. Isaieva, PA Vuissoz, and J. Leclere. Modeling the temporal evolution of the vocal tract shape with deep learning techniques. In *Procedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, 2023.

[22] Vinicius Ribeiro, Yiteng Huang, Yuan Shangguan, Zhaojun Yang, Li Wan, and Ming Sun. Handling the Alignment for Wake Word Detection: A Comparison Between Alignment-Based, Alignment-Free and Hybrid Approaches. In *Proc. INTERSPEECH 2023*, pages 5366–5370, 2023.

[23] Romain Karpinski, Vinicius Ribeiro, and Yves Laprie. Accelerating the centerline processing of vocal tract shapes for articulatory synthesis. In *Internationl Congress of Acoustics*, 2022.

[24] Wikipedia Commons. Orgue de l'église saint germain l'auxerrois - paris, 2007. URL https://commons.wikimedia.org/wiki/File:StGermainAuxerrois1.jpg. [Online; accessed 11-Aug-2023].

[25] William Henry Stone. *Elementary Lessons on Sound*. MacMillan & Co, 1879.

[26] Ingo R Titze. The human instrument. *Scientific American*, 298(1):94–101, 2008.

[27] Ingo R Titze and Daniel W Martin. Principles of voice production, 1998.

[28] Wikipedia Commons. Medical Gallery of Blausen Medical 2014, 2014. URL https://commons.wikimedia.org/wiki/File:Mouth_and_pharynx.png. [Online; accessed 15-May-2023].

[29] Reed Blaylock, Louis Goldstein, and Shrikanth S. Narayanan. Velum Control for Oral Sounds. In *Proc. INTERSPEECH 2016*, pages 1084–1088, 2016.

[30] Takatoshi Iida, Haruka Tohara, Satoko Wada, Ayako Nakane, Ryuichi Sanpei, and Koichiro Ueda. Aging decreases the strength of suprahyoid muscles involved in swallowing movements. *The Tohoku journal of experimental medicine*, 231(3):223–228, 2013.

[31] Shinji Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling*, pages 131–149. Springer, 1990.

[32] Louis-Jean Boë, Jean-Louis Heim, Kiyoshi Honda, Shinji Maeda, Pierre Badin, and Christian Abry. The vocal tract of newborn humans and neanderthals: Acoustic capabilities and consequences for the debate on the origin of language. a reply to lieberman (2007a). *Journal of Phonetics*, 35(4):564–581, 2007.

[33] I. Douros, J. Felblinger, J. Frahm, K. Isaieva, A. Joseph, Y. Laprie, F. Odille, A. Tsukanova, D. Voit, and PA Vuissoz. A multimodal real-time MRI articulatory corpus of french for speech research. In *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.

[34] James L Hiatt. The oral cavity, palate, and pharynx. In *Textbook of head and neck anatomy*, chapter 4. Jones & Bartlett Learning, 2020.

[35] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg. Silent speech interfaces. *Speech Communication*, 52(4):270–287, 2010.

[36] Gilbert Brock and Nicolas Tassi. fruleux-42.flv, 2021.

[37] Gilbert Brock and Nicolas Tassi. hirsch_iksky_vr.flv, 2021.

[38] Gilbert Brock and Nicolas Tassi. sock_ata_vr.flv, 2021.

[39] R Sock, F Hirsch, Y Laprie, P Perrier, B Vaxelaire, G Brock, F Bouarourou, C Fauth, V Hecker, L Ma, et al. Docvacim an x-ray database and tools for the study of coarticulation, inversion and evaluation of physical models. In *The Ninth International Seminar on Speech Production-ISSP*, volume 11, pages 41–48, 2011.

[40] Kenneth L Moll. Cinefluorographic techniques in speech research. *Journal of Speech and Hearing Research*, 3(3):227–241, 1960.

[41] Shigeru Kiritani. X-ray microbeam method for measurement of articulatory dynamics-techniques and results. *Speech Communication*, 5(2):119–140, 1986. ISSN 0167-6393. Speech Research in Japan.

[42] Valentina S Lucas, Ruth S Burk, Sue Creehan, and Mary Jo Grap. Utility of high-frequency ultrasound: moving beyond the surface to detect changes in skin integrity. *Plastic surgical nursing: official journal of the American Society of Plastic and Reconstructive Surgical Nurses*, 34(1):34, 2014.

[43] D Jasaitiene, S Valiukeviciene, G Linkeviciute, R Raisutis, E Jasiuniene, and R Kazys. Principles of high-frequency ultrasonography for investigation of skin pathology. *Journal of the European Academy of Dermatology and Venereology*, 25(4):375–382, 2011.

[44] Charles A Kelsey, Fred D Minifie, and Thomas J Hixon. Applications of ultrasound in speech research. *Journal of Speech and Hearing Research*, 12(3):564–575, 1969.

[45] Thomas H Shawker, Barbara Sonies, Maureen Stone, and Bruce J Baum. Real-time ultrasound visualization of tongue movement during swallowing. *Journal of Clinical Ultrasound*, 11(9):485–490, 1983.

[46] James F Bosma, Lorna G Hepburn, Stuart D Josell, and Kathleen Baker. Ultrasound demonstration of tongue motions during suckle feeding. *Developmental Medicine & Child Neurology*, 32(3):223–229, 1990.

[47] Natalia Zharkova. Using ultrasound to quantify tongue shape and movement characteristics. *The cleft palate-craniofacial journal*, 50(1):76–81, 2013.

[48] Maureen Stone. A guide to analysing tongue motion from ultrasound images. *Clinical linguistics & phonetics*, 19(6-7):455–501, 2005.

[49] Fred D Minifie, Charles A Kelsey, James A Zagzebski, and Thomas W King. Ultrasonic scans of the dorsal surface of the tongue. *The Journal of the Acoustical Society of America*, 49(6B):1857–1860, 1971.

[50] Avraham Parush, David J Ostry, and Kevin G Munhall. A kinematic study of lingual coarticulation in VCV sequences. *The Journal of the Acoustical Society of America*, 74(4):1115–1125, 1983.

[51] Eric Keller. Ultrasound measurements of tongue dorsum movements in articulatory speech impairments. *Phonetic approaches to speech production in aphasia and related disorders*, pages 93–112, 1987.

[52] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, 52(4):288–300, 2010. ISSN 0167-6393. Silent Speech Interfaces.

[53] Tamás Gábor Csapó, Csaba Zainkó, László Tóth, Gábor Gosztolya, and Alexandra Markó. Ultrasound-Based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis. pages 2727–2731, 2020.

[54] Joseph S Perkell, Marc H Cohen, Mario A Svirsky, Melanie L Matthies, Iñaki Garabieta, and Michel TT Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, 92(6):3078–3096, 1992.

[55] P. Branderud. Movetrack – a movement tracking system. In *Proceedings of the French-Swedish Symposium on Speech, Grenoble, France*, 1982.

[56] Bernd J Kröger, Marianne Pouplier, and Mark K Tiede. An evaluation of the aurora system as a flesh-point tracking tool for speech production research. 2008.

[57] Christopher Dromey, Elise Hunter, and Shawn L Nissen. Speech adaptation to kinematic recording sensors: Perceptual and acoustic findings. *Journal of Speech, Language, and Hearing Research*, 61(3):593–603, 2018.

[58] Teja Rebernik, Jidde Jacobi, Roel Jonkers, Aude Noiray, and Martijn Wieling. A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, 12(1), 2021.

[59] S Smith and R Aasen. The effects of electromagnetic fields on cardiac pacemakers. *IEEE Transactions on Broadcasting*, 38(2):136–139, 1992.

[60] Gabriella Tognola, Marta Parazzini, Federica Sibella, Alessia Paglialonga, and Paolo Ravazzani. Electromagnetic interference and cochlear implants. *Annali dell'Istituto superiore di sanita*, 43(3):241–247, 2007.

[61] Abiodun M Aibinu, Momoh-Jimoh E Salami, Amir A Shafie, and Athaur R Najeeb. MRI reconstruction using discrete fourier transform: a tutorial. 2008.

[62] Wikipedia Commons. Human anatomy planes, 2008. URL https://commons.wikimedia.org/wiki/File:Human_anatomy_planes.svg. [Online; accessed 15-May-2023].

[63] J Rokkaku. Measurements of the three-dimensional shape of the vocal tract based on the magnetic resonance imaging technique. *Ann. Bull. RILP*, 20:47–54, 1986.

[64] T Baer, JC Gore, S Boyce, and PW Nye. Application of MRI to the analysis of speech production. *Magnetic resonance imaging*, 5(1):1–7, 1987.

[65] Martin Uecker, Shuo Zhang, Dirk Voit, Alexander Karaus, Klaus-Dietmar Merboldt, and Jens Frahm. Real-time MRI at a resolution of 20 ms. *NMR in Biomedicine*, 23(8):986–994, 2010.

[66] Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd. An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America*, 115(4):1771–1776, 2004.

[67] Olov Engwall. Assessing MRI measurements: Effects of sustenation, gravitation and coarticulation. 2006.

[68] Yinghua Zhu, Asterios Toutios, Shrikanth S Narayanan, and Krishna S Nayak. Faster 3D vocal tract real-time MRI using constrained reconstruction. In *Proc. INTERSPEECH 2013*, pages 1292–1296, 2013.

[69] Ziwei Zhao, Yongwan Lim, Dani Byrd, Shrikanth Narayanan, and Krishna S Nayak. Improved 3d real-time mri of speech production. *Magnetic Resonance in Medicine*, 85(6):3182–3195, 2021.

[70] Donald G. Miller, Arend M. Sulter, Harm K. Schutte, and Rienhart F. Wolf. Comparison of vocal tract formants in singing andnonperiodic phonation. *Journal of Voice*, 11(1):1–11, 1997.

[71] Peter W. Iltis, Jens Frahm, Dirk Voit, Arun A. Joseph, Erwin Schoonderwaldt, and Eckart Altenmüller. High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players. *Quantitative imaging in medicine and surgery*, 5(3):374, 2015.

[72] Michael Proctor, Erik Bresch, Dani Byrd, Krishna Nayak, and Shrikanth Narayanan. Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 133(2):1043–1054, 2013.

[73] Reed Blaylock and Shrikanth Narayanan. Beatboxing Kick Drum Kinematics. In *Proc. INTERSPEECH 2023*, pages 2583–2587, 2023.

[74] Sajan Goud Lingala, Brad P Sutton, Marc E Miquel, and Krishna S Nayak. Recommendations for real-time speech mri. *Journal of Magnetic Resonance Imaging*, 43(1):28–44, 2016.

[75] E. Bellon, M. Haacke, P. Coleman, D. Sacco, DA Steiger, and R. Gangarosa. MR artifacts: A review. *AJR. American journal of roentgenology*, 147:1271–81, 12 1986. doi: 10.2214/ajr.147.6.1271.

[76] K. G. Munhall, E. Vatikiotis-Bateson, and Y. Tohkura. X-ray film database for speech research. *The Journal of the Acoustical Society of America*, 98(2):1222–1224, 08 1995. ISSN 0001-4966. doi: 10.1121/1.413621.

[77] John R Westbury, Greg Turner, and J Dembowski. X-ray microbeam speech production database user's handbook. *University of Wisconsin*, 1994.

[78] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A Bilmes. Unsupervised learning of acoustic features via deep canonical correlation analysis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4590–4594. IEEE, 2015.

[79] Raman Arora and Karen Livescu. Multi-view learning with supervision for transformed bottleneck features. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2499–2503. IEEE, 2014.

[80] Rudolph Sock, Fabrice Hirsch, Yves Laprie, Pascal Perrier, Béatrice Vaxelaire, Gilbert Brock, Fayssal Bouarourou, Camille Fauth, Véronique Ferbach-Hecker, Liang Ma, et al. An x-ray database, tools and procedures for the study of speech production. In *ISSP 2011-9th International Seminar on Speech Production*, pages 41–48, 2011.

[81] Alan Wrench. The MOCHA-TIMIT articulatory database, 1999.

[82] K. Richmond, P. Hoole, and S. King. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[83] Ingmar Steiner, Korin Richmond, Ian Marshall, and Calum D Gray. The magnetic resonance imaging subset of the mngu0 articulatory corpus. *The Journal of the Acoustical Society of America*, 131(2):EL106–EL111, 2012.

[84] Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, et al. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). *The Journal of the Acoustical Society of America*, 136(3):1307–1311, 2014.

[85] Tanner Sorensen, Zisis Iason Skordilis, Asterios Toutios, Yoon-Chul Kim, Yinghua Zhu, Jangwon Kim, Adam C Lammert, Vikram Ramanarayanan, Louis Goldstein, Dani Byrd, et al. Database of volumetric and real-time vocal tract MRI for speech science. In *Proc. INTERSPEECH 2017*, pages 645–649, 2017.

[86] António Teixeira, Paula Martins, Catarina Oliveira, Carlos Ferreira, Augusto Silva, and Ryan Shosted. Real-time mri for portuguese: database, methods and applications. In *Computational Processing of the Portuguese Language: 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings 10*, pages 306–317. Springer, 2012.

[87] Karyna Isaieva, Yves Laprie, Justine Leclère, Ioannis K Douros, Jacques Felblinger, and Pierre-André Vuissoz. Multimodal dataset of real-time 2D and static 3D MRI of healthy french speakers. *Scientific Data*, 8(1):258, 2021.

[88] Yongwan Lim, Asterios Toutios, Yannick Bliesener, Ye Tian, Sajan Goud Lingala, Colin Vaz, Tanner Sorensen, Miran Oh, Sarah Harper, Weiyi Chen, et al. A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images. *Scientific data*, 8(1):1–14, 2021.

[89] Jangwon Kim, Asterios Toutios, Yoon-Chul Kim, Yinghua Zhu, Sungbok Lee, and Shrikanth Narayanan. USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging. In *International Seminar on Speech Production (ISSP), Cologne, Germany*, volume 226, 2014.

[90] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[91] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[92] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[93] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[94] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[95] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[96] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[97] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[98] Yann LeCun. The MNIST database of handwritten digits. 1998.

[99] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[100] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[101] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[102] Donald O Hebb. Organization of behavior: A neurophysiological theory. *(No Title)*, 1949.

[103] Bhiksha Raj. Lecture notes in introduction to neural networks, 2021. URL http://lxmls.it.pt/2021/wp-content/uploads/2021/07/bhiksha-lecture.pdf.

[104] Marvin Minsky and Seymour Papert. An introduction to computational geometry. *Cambridge tiass., HIT*, 479:480, 1969.

[105] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer, 2015.

[106] Simeon Kostadinov. Understanding backpropagation algorithm, 2019. URL https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd.

[107] Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.

[108] Leslie N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

[109] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[110] Yan LeCun. *Connectionist Learning Models*. PhD thesis, Université Pierre et Marie Currie, Paris, France, 1987.

[111] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.

[112] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6, 1993.

[113] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv preprint arXiv:2003.05991*, 2020.

[114] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[115] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[116] Saïd Ladjal, Alasdair Newson, and Chi-Hieu Pham. A PCA-like autoencoder. *arXiv preprint arXiv:1904.01277*, 2019.

[117] Naohiro Tawara, Tetsunori Kobayashi, and Tetsuji Ogawa. Multi-channel speech enhancement using time-domain convolutional denoising autoencoder. In *Proc. INTERSPEECH 2019*, pages 86–90, 2019.

[118] Walid El-Shafai, SA El-Nabi, E El-Rabaie, A Ali, F Soliman, Abeer D Algarni, and FEA El-Samie. Efficient deep-learning-based autoencoder denoising approach for medical image diagnosis. *Comput. Mater. Contin*, 70:6107–6125, 2022.

[119] Peng Liang, Wenzhong Shi, and Xiaokang Zhang. Remote sensing image classification based on stacked denoising autoencoder. *Remote Sensing*, 10(1):16, 2017.

[120] Xinya Wu, Yan Zhang, Changming Cheng, and Zhike Peng. A hybrid classification autoencoder for semi-supervised fault diagnosis in rotating machinery. *Mechanical Systems and Signal Processing*, 149:107327, 2021.

[121] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. Deep clustering with convolutional autoencoders. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24*, pages 373–382. Springer, 2017.

[122] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4066–4075, 2019.

[123] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674, 2017.

[124] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[125] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. Variational autoencoder for semi-supervised text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[126] Xiangru Chen, Yanan Sun, Mengjie Zhang, and Dezhong Peng. Evolving deep convolutional variational autoencoders for image classification. *IEEE Transactions on Evolutionary Computation*, 25(5):815–829, 2020.

[127] Ertuğ Karamatlı, Ali Taylan Cemgil, and Serap Kırbız. Audio source separation using variational autoencoders and weak class supervision. *IEEE Signal Processing Letters*, 26(9):1349–1353, 2019.

[128] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[129] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[130] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[131] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Learning what and where to transfer. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3030–3039. PMLR, 09–15 Jun 2019.

[132] Afonso Menegola, Michel Fornaciali, Ramon Pires, Sandra Avila, and Eduardo Valle. Towards automated melanoma screening: Exploring transfer learning schemes. *arXiv preprint arXiv:1609.01228*, 2016.

[133] Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. Transfer learning for speech recognition on a budget. *arXiv preprint arXiv:1706.00290*, 2017.

[134] Dong Wang and Thomas Fang Zheng. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237. IEEE, 2015.

[135] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[136] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[137] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022.

[138] Abdolreza Sabzi Shahrebabaki, Negar Olfati, Sabato Marco Siniscalchi, Giampiero Salvi, and Torbjørn Svendsen. Transfer learning of articulatory information through phone information. In *INTERSPEECH*, pages 2877–2881, 2020.

[139] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[140] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017.

[141] Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. Self-supervised knowledge distillation using singular value decomposition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 335–350, 2018.

[142] Chenrui Zhang and Yuxin Peng. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. *arXiv preprint arXiv:1804.10069*, 2018.

[143] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[144] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[145] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[146] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[147] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[148] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[149] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[150] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[151] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[152] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[153] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[154] Alceu Bissoto, Fábio Perez, Vinícius Ribeiro, Michel Fornaciali, Sandra Avila, and Eduardo Valle. Deep-learning ensembles for skin-lesion segmentation, analysis, classification: Recod titans at isic challenge 2018. *arXiv preprint arXiv:1808.08480*, 2018.

[155] Eduardo Valle, Michel Fornaciali, Afonso Menegola, Julia Tavares, Flávia Vasques Bittencourt, Lin Tzy Li, and Sandra Avila. Data, depth, and design: Learning reliable models for skin lesion analysis. *Neurocomputing*, 383:303–313, 2020.

[156] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[157] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

[158] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[159] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[160] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[161] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[162] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019.

[163] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[164] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[165] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012.

[166] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[167] Getao Du, Xu Cao, Jimin Liang, Xueli Chen, and Yonghua Zhan. Medical image segmentation based on u-net: A review. *Journal of Imaging Science and Technology*, 2020.

[168] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[169] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

[170] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[171] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[172] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. Less is more: Sample selection and label conditioning improve skin lesion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 738–739, 2020.

[173] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[174] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[175] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[176] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[177] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[178] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. Handling inter-annotator agreement for automated skin lesion segmentation. *arXiv preprint arXiv:1906.02415*, 2019.

[179] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[180] Yujian Mo, Yan Wu, Xinneng Yang, Feilin Liu, and Yujun Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646, 2022.

[181] Monica Franzese and Antonella Iuliano. Hidden markov models. In Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 753–762. Academic Press, Oxford, 2019. ISBN 978-0-12-811432-2.

[182] Wai-Ki Ching and Michael K Ng. Markov chains. *Models, algorithms and applications*, 2006.

[183] Lawrence Rabiner and Biinghwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[184] Ramaprasad Bhar and Shigeyuki Hamori. *Hidden Markov models: applications to financial economics*, volume 40. Springer Science & Business Media, 2004.

[185] Matthew S Crouse, Robert D Nowak, and Richard G Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on signal processing*, 46(4):886–902, 1998.

[186] Antti Koski. Modelling ecg signals with hidden markov models. *Artificial intelligence in medicine*, 8(5): 453–471, 1996.

[187] Mark Gales, Steve Young, et al. The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304, 2008.

[188] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252, 2013.

[189] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[190] Michael I Jordan. Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pages 471–495. Elsevier, 1997.

[191] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[192] Ronald J Williams and David Zipser. *Gradient-based learning algorithms for recurrent connectionist networks*. Citeseer, 1990.

[193] Ilya Sutskever. *Training recurrent neural networks*. University of Toronto Toronto, ON, Canada, 2013.

[194] James Martens and Ilya Sutskever. Learning recurrent neural networks with hessian-free optimization. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1033–1040, 2011.

[195] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

[196] Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8624–8628. IEEE, 2013.

[197] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[198] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[199] Felix A Gers and E Schmidhuber. Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE transactions on neural networks*, 12(6):1333–1340, 2001.

[200] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[201] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[202] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[203] Johannes Maucher. Sequence-to-sequence, attention, transformer, 2021. URL https://hannibunny.github.io/mlbook/transformer/attention.html.

[204] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[205] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28, 2015.

[206] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[207] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[208] Mary Phuong and Marcus Hutter. Formal algorithms for transformers. *arXiv preprint arXiv:2207.09238*, 2022.

[209] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[210] Nicole Holliday. Siri, you've changed! acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication*, 8:1116955, 2023.

[211] Peter Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8(4): e60603, 2013.

[212] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*, 2023.

[213] Linda I Shuster, Dennis M Ruscello, and Kimberly D Smith. Evoking [r] using visual feedback. *American Journal of Speech-Language Pathology*, 1(3):29–34, 1992.

[214] Fiona E Gibbon and Sara E Wood. Visual feedback therapy with electropalatography. In *Interventions in speech sound disorders*, pages 509–536. Paul H. Brookes Publishing Co., Inc., 2010.

[215] Penelope Bacsfalvi and Barbara May Bernhardt. Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: Ultrasound and electropalatography. *Clinical Linguistics & Phonetics*, 25(11-12):1034–1043, 2011.

[216] Atsuo Suemitsu, Jianwu Dang, Takayuki Ito, and Mark Tiede. A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning. *The Journal of the Acoustical Society of America*, 138(4):EL382–EL387, 2015.

[217] June S Levitt and William F Katz. Augmented visual feedback in second language learning: Training japanese post-alveolar flaps to american english speakers. In *Proceedings of Meetings on Acoustics 154ASA*, volume 2, page 060002. Acoustical Society of America, 2007.

[218] Olov Engwall. Can audio-visual instructions help learners improve their articulation?-an ultrasound study of short term changes. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[219] Pramit Saha, Praneeth Srungarapu, and Sidney Fels. Towards automatic speech identification from vocal tract shape dynamics in real-time MRI. *arXiv preprint arXiv:1807.11089*, 2018.

[220] Sven EG Öhman. Coarticulation in vcv utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39(1):151–168, 1966.

[221] Jonas Beskow. Trainable articulatory control models for visual speech synthesis. *International Journal of Speech Technology*, 7(4):335–349, 2004.

[222] Michael M Cohen and Dominic W Massaro. Modeling coarticulation in synthetic visual speech. In *Models and techniques in computer animation*, pages 139–156. Springer, 1993.

[223] Sven EG Öhman. Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, 41(2):310–320, 1967.

[224] Anders Löfqvist. Speech as audible gestures. In *Speech production and speech modelling*, pages 289–322. Springer, 1990.

[225] C. H. Coker. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64(4):452–460, 1976.

[226] S Maeda. Un modele articulatoire de la langue avec des composantes lineaires. 10eme journées d'etude sur la parole, 1979.

[227] Mazin G Rahim, Colin C Goodyear, W Bastiaan Kleijn, Juergen Schroeter, and M Mohan Sondhi. On the use of neural networks in articulatory speech synthesis. *The Journal of the Acoustical Society of America*, 93(2):1109–1121, 1993.

[228] Victor N Sorokin, Alexander S Leonov, and Alexander V Trushkin. Estimation of stability and accuracy of inverse problem solution for the vocal tract. *Speech Communication*, 30(1):55–74, 2000.

[229] B. Potard, Y. Laprie, and S. Ouni. Incorporation of phonetic constraints in acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 123(4):2310–2323, 2008.

[230] Peter Birkholz and Dietmar Jackel. A three-dimensional model of the vocal tract for speech synthesis. In *Proceedings of the 15th international congress of phonetic sciences*, pages 2597–2600. Barcelona, Spain, 2003.

[231] Benjamin Elie and Yves Laprie. Copy synthesis of running speech based on vocal tract imaging and audio recording. In *22nd International Congress on Acoustics (ICA)*, 2016.

[232] K Ishizaka and JL Flanagan. Acoustic properties of a two-mass model of the vocal cords. *The Journal of the Acoustical Society of America*, 51(1A):91–91, 1972.

[233] Benjamin Elie and Yves Laprie. Acoustic impact of the gradual glottal abduction degree on the production of fricatives: A numerical study. *The Journal of the Acoustical Society of America*, 142(3):1303–1317, 2017.

[234] C. P. Browman and L. Goldstein. Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180, 1992.

[235] H. Nam, V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Espy-Wilson, and M. Hasegawa-Johnson. A procedure for estimating gestural scores from natural speech. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[236] P. Birkholz, B. J. Kroger, and C. Neuschaefer-Rube. Model-based reproduction of articulatory trajectories for consonant–vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5): 1422–1433, 2010.

[237] Benjamin Elie, Juraj Simko, and Alice Turk. Optimal control of speech with context-dependent articulatory targets. In *Proc. INTERSPEECH 2023*. ISCA, 2023.

[238] Alice Turk and Stefanie Shattuck-Hufnagel. *Speech timing: Implications for theories of phonology, phonetics, and speech motor control*, volume 5. Oxford University Press, USA, 2020.

[239] K. Richmond. Preliminary inversion mapping results with a new ema corpus. 2009.

[240] Christopher M Bishop. Mixture density networks. 1994.

[241] T. Biasutto-Lervat and S. Ouni. Phoneme-to-articulatory mapping using bidirectional gated rnn. In *Proc. INTERSPEECH 2018*, 2018.

[242] T. Biasutto, S. Dahmani, S. Ouni, et al. Modeling labial coarticulation with bidirectional gated recurrent networks and transfer learning. In *Proc. INTERSPEECH 2019*, 2019.

[243] Tamás Gábor Csapó. Speaker dependent acoustic-to-articulatory inversion using real-time MRI of the vocal tract. *arXiv preprint arXiv:2008.02098*, 2020.

[244] Gábor Gosztolya, Ádám Pintér, László Tóth, Tamás Grósz, Alexandra Markó, and Tamás Gábor Csapó. Autoencoder-based articulatory-to-acoustic mapping for ultrasound silent speech interfaces. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[245] Ahmed Adel Attia and Carol Y Espy-Wilson. Masked autoencoders are articulatory learners. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[246] Gunnar Fant. Acoustic theory of speech production. Mouton, The Hague, 1960.

[247] Yves Laprie, Matthieu Loosvelt, Shinji Maeda, Rudolph Sock, and Fabrice Hirsch. Articulatory copy synthesis from cine x-ray films. In *Proc. INTERSPEECH 2013*, pages 2024–2028, 2013.

[248] Shinji Maeda. A digital simulation method of the vocal-tract system. *Speech communication*, 1(3-4):199–229, 1982.

[249] John M Heinz and Kenneth N Stevens. On the derivation of area functions and acoustic spectra from cineradiographic films of speech. *The Journal of the Acoustical Society of America*, 36(5_Supplement): 1037–1038, 1964.

[250] Anton A Poznyakovskiy, Alexander Mainka, Ivan Platzek, Dirk Mürbe, et al. A fast semiautomatic algorithm for centerline-based vocal tract segmentation. *BioMed Research International*, 2015, 2016.

[251] Zisis Iason Skordilis, Asterios Toutios, Johannes Töger, and Shrikanth Narayanan. Estimation of vocal tract area function from volumetric magnetic resonance imaging. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 924–928. IEEE, 2017.

[252] Paul Konstantin Krug, Simon Stone, and Peter Birkholz. Intelligibility and naturalness of articulatory synthesis with vocaltractlab compared to established speech synthesis technologies. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 102–107, 2021.

[253] Daniel R van Niekerk, Anqi Xu, Branislav Gerazov, Paul K Krug, Peter Birkholz, Lorna Halliday, Santitham Prom-on, and Yi Xu. Simulating vocal learning of spoken language: Beyond imitation. *Speech Communication*, 147:51–62, 2023.

[254] Thanat Lapthawan, Santitham Prom-On, Peter Birkholz, and Yi Xu. Estimating underlying articulatory targets of thai vowels by using deep learning based on generating synthetic samples from a 3d vocal tract model and data augmentation. *IEEE Access*, 10:41489–41502, 2022.

[255] Yingming Gao, Simon Stone, and Peter Birkholz. Articulatory Copy Synthesis Based on a Genetic Algorithm. In *Proc. INTERSPEECH 2019*, pages 3770–3774, 2019.

[256] Beiming Cao, Alan Wisler, and Jun Wang. Speaker adaptation on articulation and acoustics for articulation-to-speech synthesis. *Sensors*, 22(16):6056, 2022.

[257] Mark Tiede, Carol Y Espy-Wilson, Dolly Goldenberg, Vikramjit Mitra, Hosung Nam, and Ganesh Sivaraman. Quantifying kinematic aspects of reduction in a contrasting rate production task. *The Journal of the Acoustical Society of America*, 141(5):3580–3580, 2017.

[258] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE, 2019.

[259] Seyed Hamidreza Mohammadi and Alexander Kain. An overview of voice conversion systems. *Speech Communication*, 88:65–82, 2017.

[260] Ian L Dryden and Kanti V Mardia. *Statistical shape analysis: with applications in R*, volume 995. John Wiley & Sons, 2016.

[261] Marc-Antoine Georges, Julien Diard, Laurent Girin, Jean-Luc Schwartz, and Thomas Hueber. Repeat after me: self-supervised learning of acoustic-to-articulatory mapping by vocal imitation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8252–8256. IEEE, 2022.

[262] Raphaël Laurent, Marie-Lou Barnaud, Jean-Luc Schwartz, Pierre Bessière, and Julien Diard. The complementary roles of auditory and motor information evaluated in a bayesian perceptuo-motor model of speech perception. *Psychological review*, 124(5):572, 2017.

[263] Antoine Serrurier, Pierre Badin, Laurent Lamalle, and Christiane Neuschaefer-Rube. Characterization of inter-speaker articulatory variability: a two-level multi-speaker modelling approach based on MRI data. *The Journal of the Acoustical Society of America*, 145(4):2149–2170, 2019.

[264] Jun Wang, Ashok Samal, and Jordan Green. Across-speaker articulatory normalization for speaker-independent silent speech recognition. 2014.

[265] Ganesh Sivaraman, Vikramjit Mitra, Hosung Nam, Mark Tiede, and Carol Espy-Wilson. Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion. *The Journal of the Acoustical Society of America*, 146(1):316–329, 2019.

[266] M. Labrunie, P. Badin, D. Voit, A. A. Joseph, J. Frahm, L. Lamalle, C. Vilain, and LJ Boë. Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. *Speech Communication*, 99:27–46, 2018.

[267] E. L. Saltzman and K. G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological psychology*, 1(4):333–382, 1989.

[268] Samuel Silva and António Teixeira. Critical Articulators Identification from RT-MRI of the Vocal Tract. In *Proc. INTERSPEECH 2017*, pages 626–630, 2017.

[269] Jangwon Kim, Asterios Toutios, Sungbok Lee, and Shrikanth S Narayanan. A kinematic study of critical and non-critical articulators in emotional speech production. *The Journal of the Acoustical Society of America*, 137(3):1411–1429, 2015.

[270] Antoine Serrurier and Christiane Neuschaefer-Rube. F1 and F2 formant variations and inter-speaker articulatory variability: A preliminary analysis. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2022*, pages 172–179, 2022.

[271] Jonathan Harrington, Steve Cassidy, Jonathan Harrington, and Steve Cassidy. The acoustic theory of speech production. *Techniques in Speech Acoustics*, pages 29–56, 1999.

[272] Robert C Streijl, Stefan Winkler, and David S Hands. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227, 2016.

[273] KG Van Leeuwen, P Bos, Stefano Trebeschi, Maarten JA van Alphen, Luuk Voskuilen, Ludi E Smeele, Ferdi van der Heijden, RJJH Van Son, et al. Cnn-based phoneme classifier from vocal tract MRI learns embedding consistent with articulatory topology. In *Proc. INTERSPEECH 2019*, pages 909–913, 2019.

[274] Olov Engwall. Evaluation of speech inversion using an articulatory classifier. In *Proc. of the Seventh International Seminar on Speech Production*, pages 431–434, 2006.

[275] Zeynab Raeesy, Sylvia Rueda, Jayaram K. Udupa, and John Coleman. Automatic segmentation of vocal tract MR images. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 1328–1331. IEEE, 2013.

[276] Sylvia Rueda and Jayaram K. Udupa. Global-to-local, shape-based, real and virtual landmarks for shape modeling by recursive boundary subdivision. In *Medical Imaging 2011: Image Processing*, volume 7962, pages 1329–1341. SPIE, 2011.

[277] Jiamin Liu and Jayaram K. Udupa. Oriented active shape models. *IEEE Transactions on medical Imaging*, 28(4):571–584, 2008.

[278] Samuel Silva and António Teixeira. Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Computer Speech & Language*, 33(1):25–46, 2015.

[279] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *European conference on computer vision*, pages 484–498. Springer, 1998.

[280] S Suganyadevi, V Seethalakshmi, and K Balasamy. A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11(1):19–38, 2022.

[281] Valliappan CA, Renuka Mannem, and Prasanta Kumar Ghosh. Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video Using Semantic Segmentation with Fully Convolutional Networks. In *Proc. INTERSPEECH 2018*, pages 3132–3136, 2018.

[282] Ian Fasel and Jeff Berry. Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. In *2010 20th International Conference on Pattern Recognition*, pages 1493–1496. IEEE, 2010.

[283] Aurore Jaumard-Hakoun, Kele Xu, Pierre Roussel-Ragot, Gérard Dreyfus, and Bruce Denby. Tongue contour extraction from ultrasound images based on deep neural network. *arXiv preprint arXiv:1605.05912*, 2016.

[284] Nan Zhang, Shifei Ding, Jian Zhang, and Yu Xue. An overview on restricted boltzmann machines. *Neurocomputing*, 275:1186–1199, 2018.

[285] Mohammad Eslami, Christiane Neuschaefer-Rube, and Antoine Serrurier. Automatic vocal tract segmentation based on conditional generative adversarial neural network. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pages 263–270, 2019.

[286] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[287] Sasan Asadiabadi and Engin Erzin. Vocal tract contour tracking in rtMRI using deep temporal regression network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3053–3064, 2020.

[288] S. Ashwin Hebbar, Rahul Sharma, Krishna Somandepalli, Asterios Toutios, and Shrikanth Narayanan. Vocal tract articulatory contour detection in real-time magnetic resonance images using spatio-temporal context. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7354–7358. IEEE, 2020.

[289] K. Isaieva, Y. Laprie, N. Turpault, A. Houssard, J. Felblinger, and PA Vuissoz. Automatic tongue delineation from MRI images with a convolutional neural network approach. *Applied Artificial Intelligence*, 34 (14):1115–1123, 2020.

[290] Yves Laprie, Benjamin Elie, Anastasiia Tsukanova, and Pierre-André Vuissoz. Centerline articulatory models of the velum and epiglottis for articulatory synthesis of speech. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2110–2114. IEEE, 2018.

[291] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[292] Edsger W Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1 (1):269–271, 1959.

[293] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[294] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and

Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[295] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[296] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[297] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep Speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[298] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[299] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[300] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[301] Linguistic Data Consortium et al. The DARPA TIMIT acoustic-phonetic continuous speech corpus. *NIST Speech CD*, pages 1–1, 1990.

[302] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[303] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460, 2020.

Appendices

Appendix A

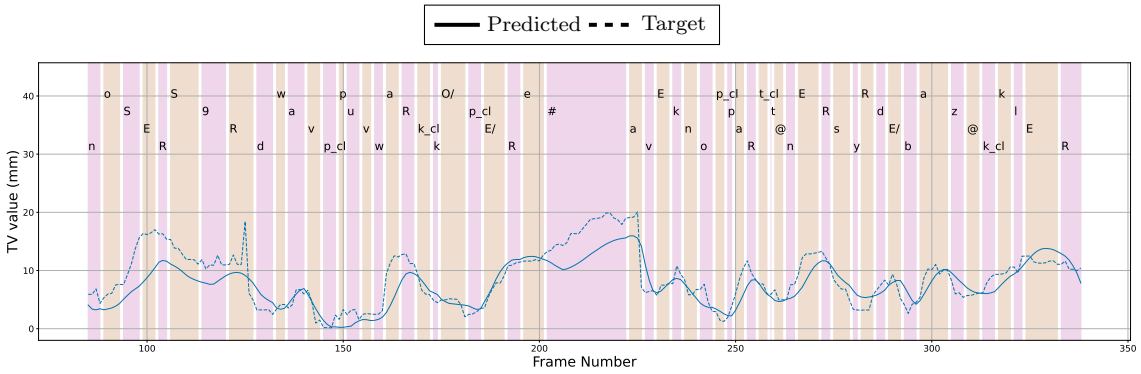
Additional Samples for Synthesized Vocal Tract Shapes

1 Model-Free Phoneme-to-Articulation

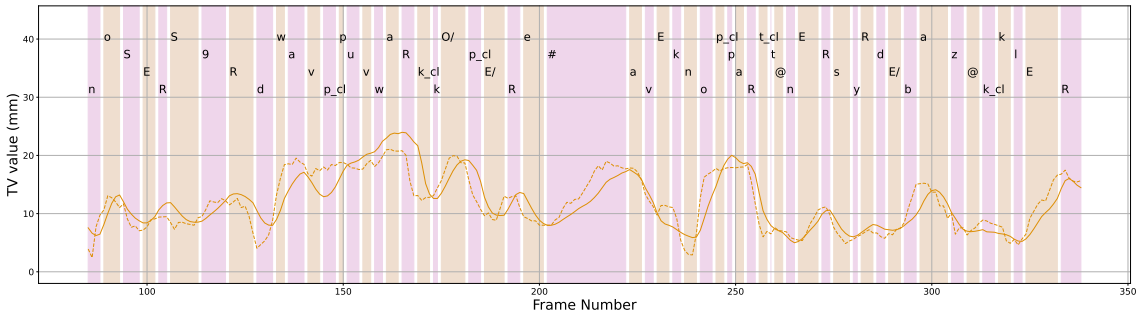
Figure A.1, Figure A.2, Figure A.3 presents the tract variables of the model-free phoneme-to-articulation predictions for three additional samples.

2 Autoencoder-Based Phoneme-to-Articulation

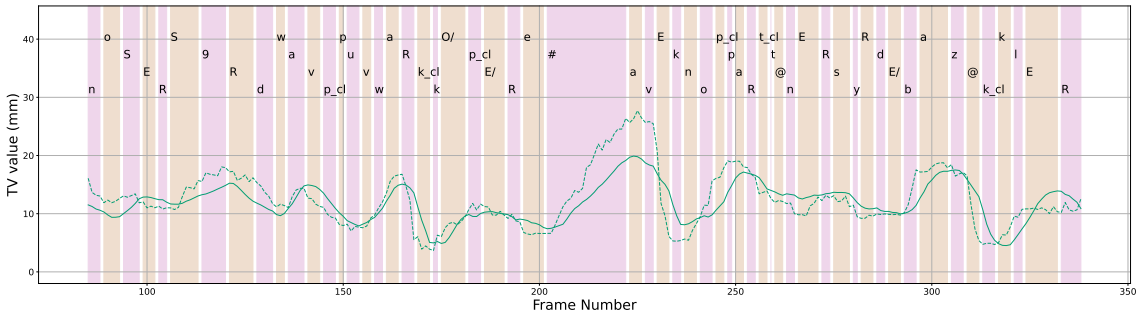
Figure A.4, Figure A.5, Figure A.6 presents the tract variables of the autoencoder-based network predictions for three additional samples in the test set.



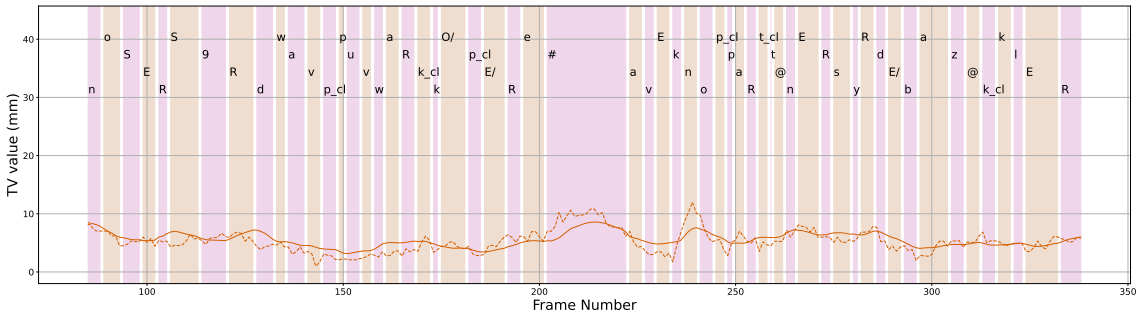
(a) LA



(b) TTCD

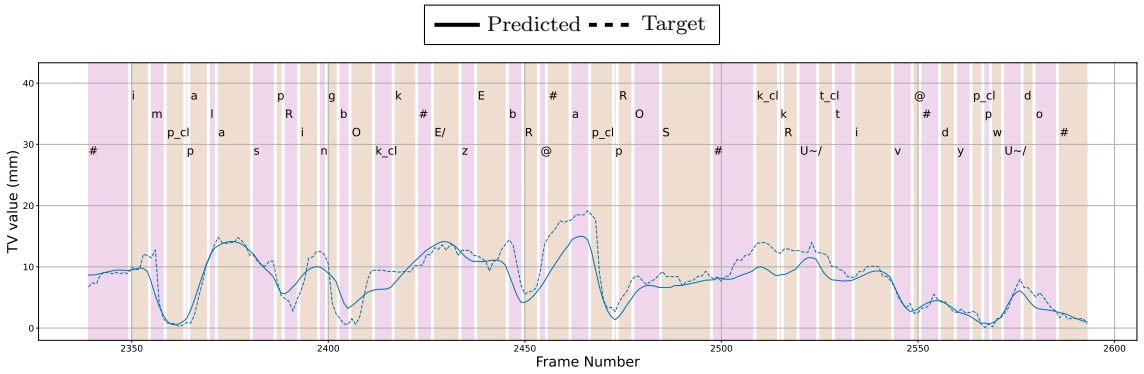


(c) TBCD

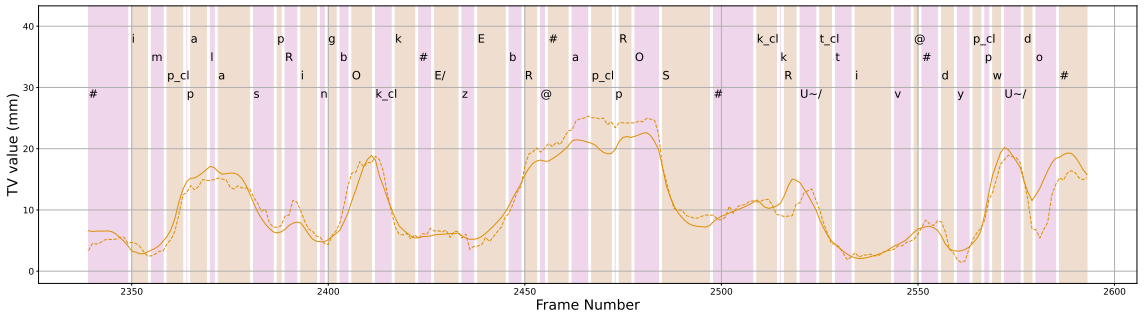


(d) VEL

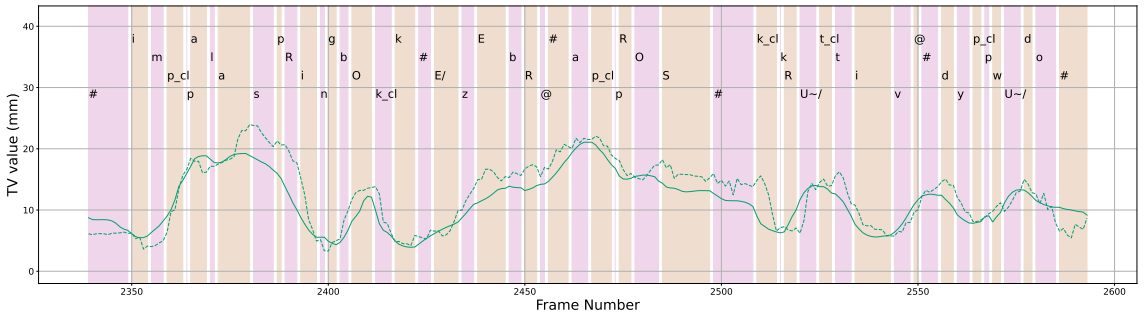
Figure A.1: Model-free phoneme-to-articulation network and ground truth TV trajectories for the utterance *"Nos chercheurs doivent pouvoir coopérer avec nos partenaires sur des bases claires."* Each image displays one tract variable. The corresponding phonemes are displayed in the top of the image. The alternating colors mark the onset and offset of each phoneme.



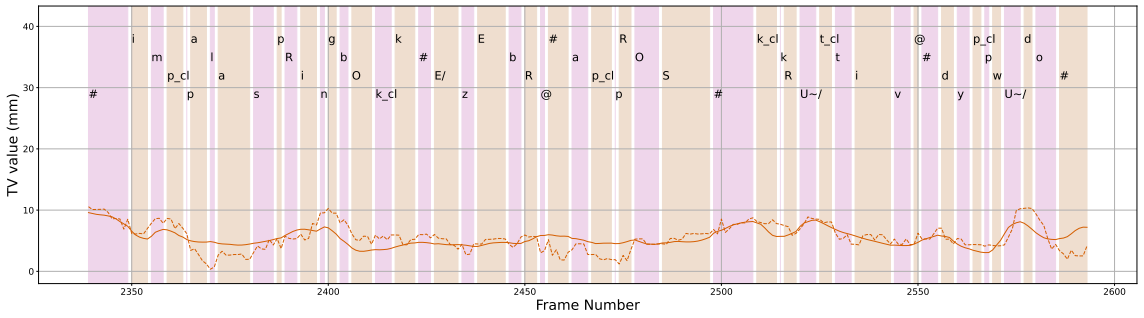
(a) LA



(b) TTCD

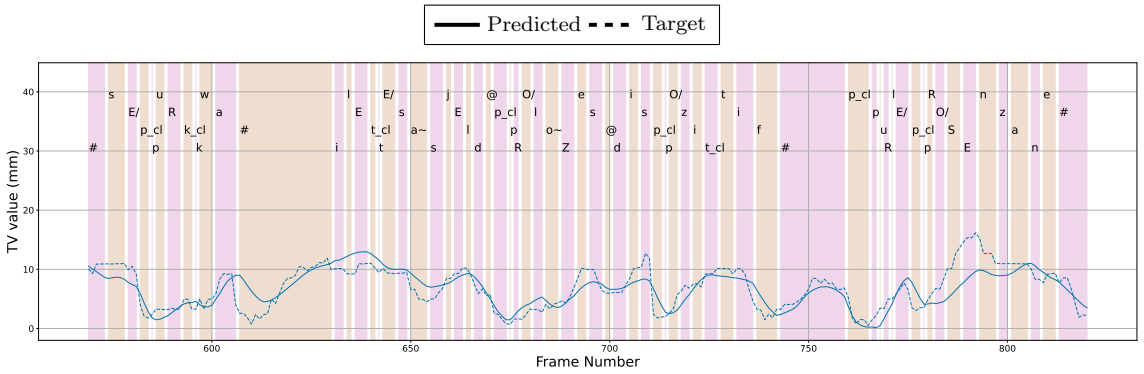


(c) TBCD

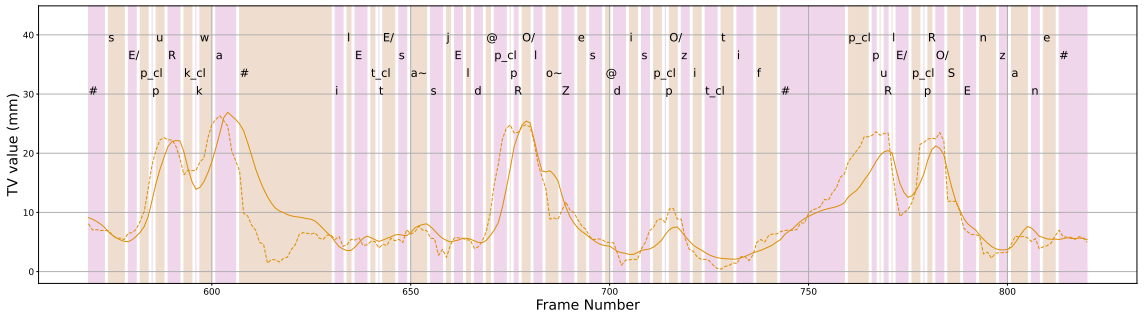


(d) VEL

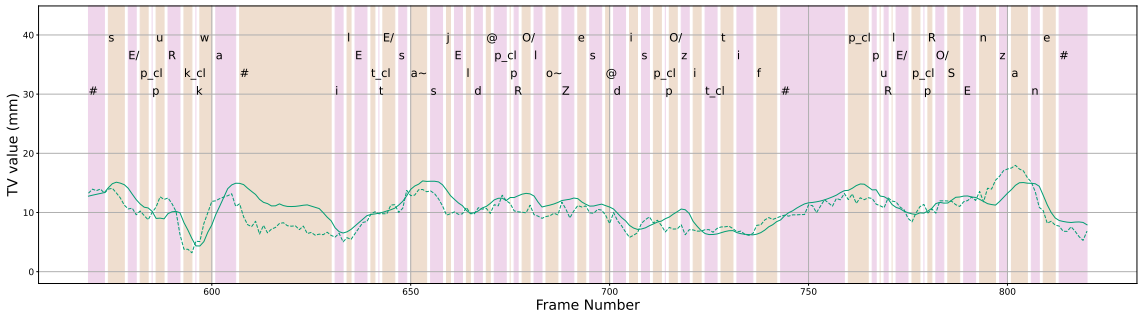
Figure A.2: Model-free phoneme-to-articulation and ground truth TV trajectories for the utterance *"Impalas, springbooks et zèbres approchent, craintives, du point d'eau."* Each image displays one tract variable. The corresponding phonemes are displayed in the top of the image. The alternating colors mark the onset and offset of each phoneme.



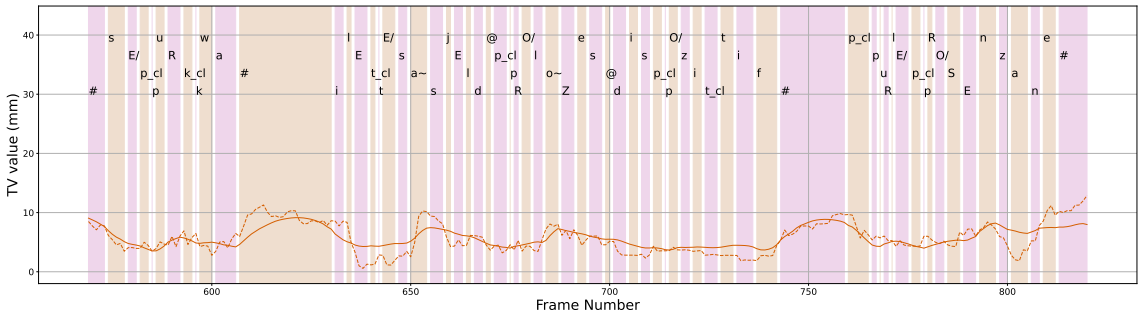
(a) LA



(b) TTCD

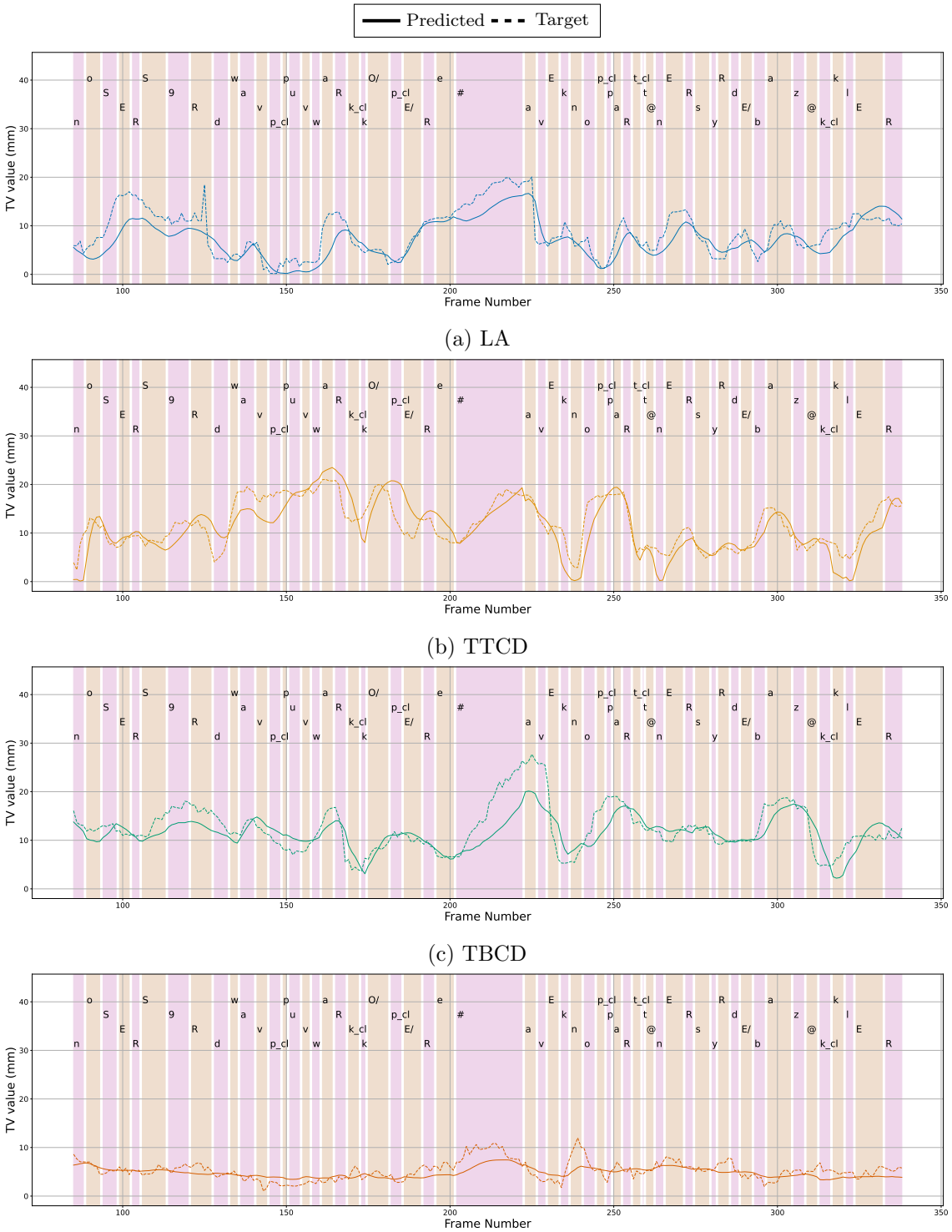


(c) TBCD



(d) VEL

Figure A.3: Model-free phoneme-to-articulation and ground truth TV trajectories for the utterance *"C'est pourquoi il est essentiel de prolonger ce dispositif pour les prochaines années."* Each image displays one tract variable. The corresponding phonemes are displayed in the top of the image. The alternating colors mark the onset and offset of each phoneme.



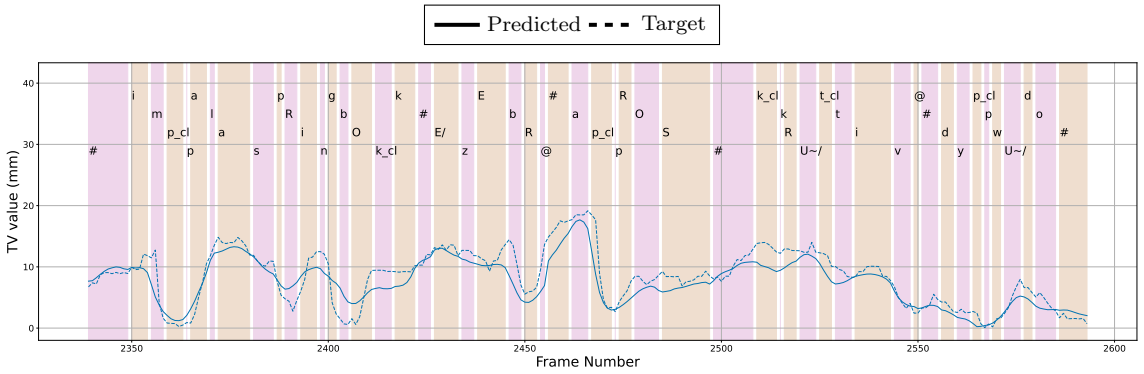
(a) LA

(b) TTCD

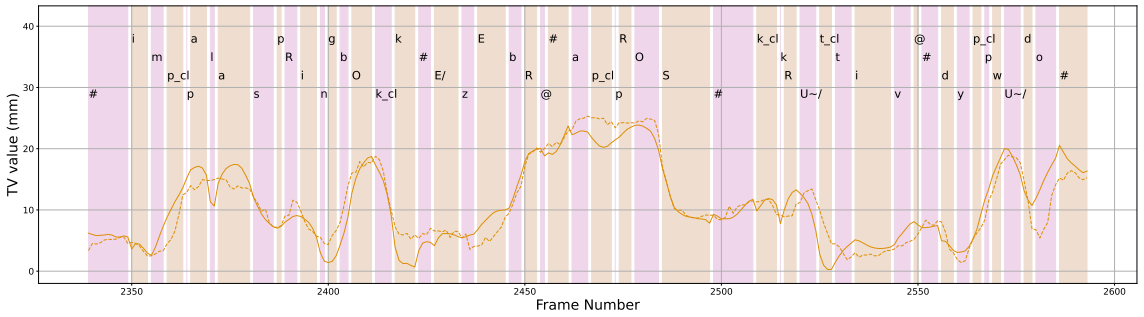
(c) TBCD

(d) VEL

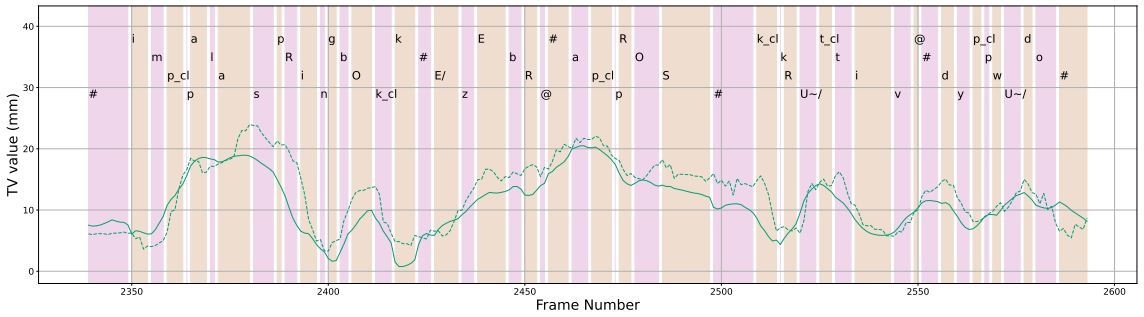
Figure A.4: Autoencoder-based phoneme-to-articulation and ground truth TV trajectories for the utterance *"Nos chercheurs doivent pouvoir coopérer avec nos partenaires sur des bases claires."* Each image displays one tract variable. The corresponding phonemes are displayed in the top of the image. The alternating colors mark the onset and offset of each phoneme.



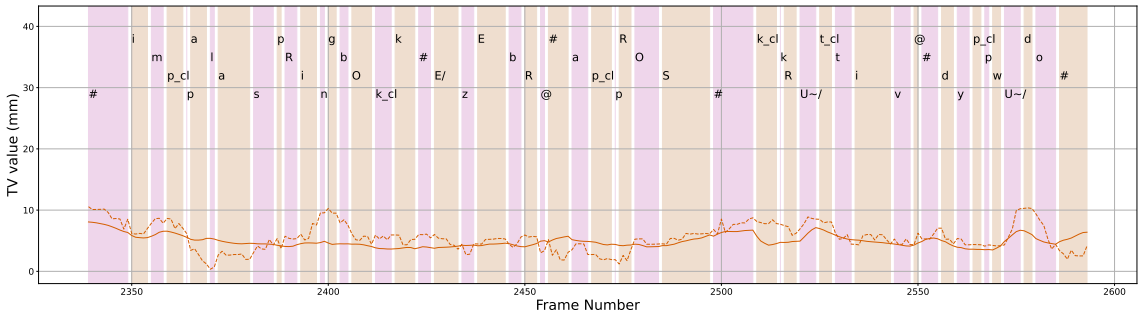
(a) LA



(b) TTCD

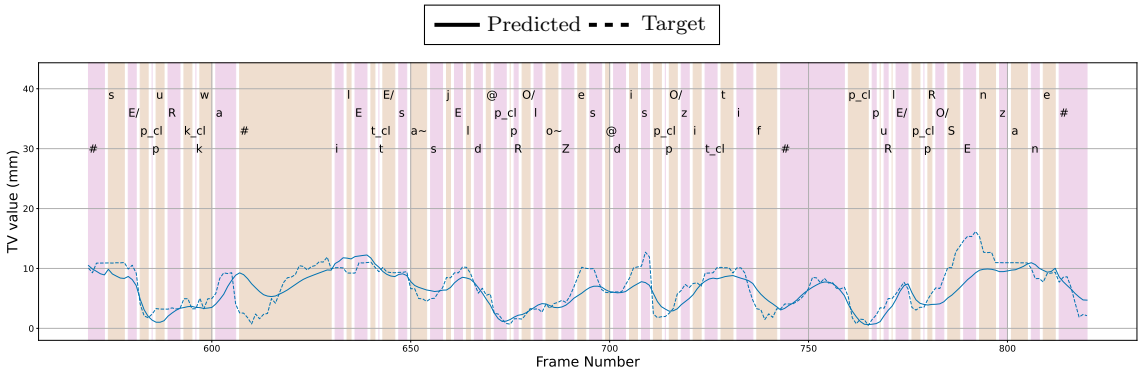


(c) TBCD

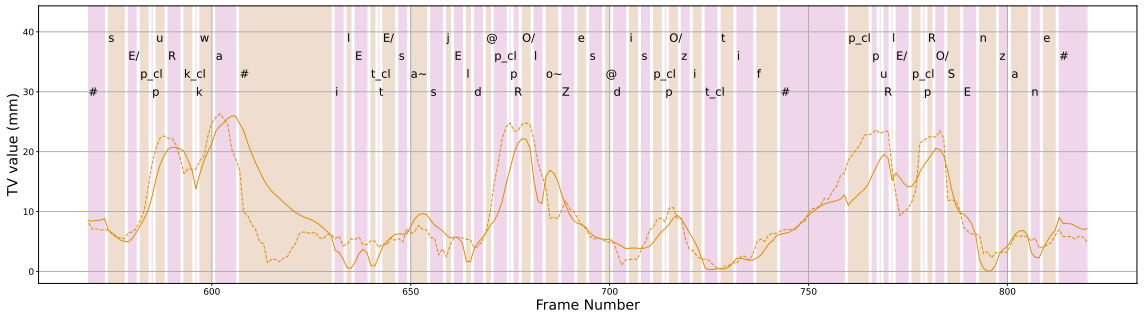


(d) VEL

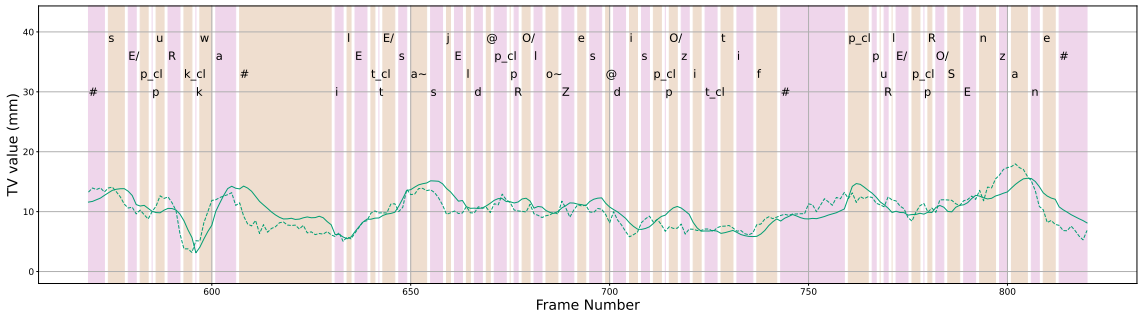
Figure A.5: Autoencoder-based phoneme-to-articulation and ground truth TV trajectories for the utterance *"Impalas, springbooks et zèbres approchent, craintives, du point d'eau."* Each image displays one tract variable. The corresponding phonemes are displayed in the top of the image. The alternating colors mark the onset and offset of each phoneme.



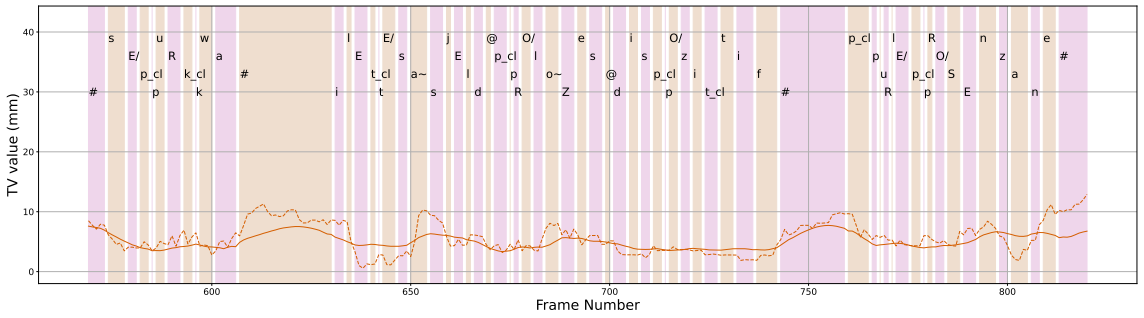
(a) LA



(b) TTCD



(c) TBCD



(d) VEL

Figure A.6: Autoencoder-based phoneme-to-articulation and ground truth TV trajectories for the utterance *"C'est pourquoi il est essentiel de prolonger ce dispositif pour les prochaines années."* Each image displays one tract variable. The corresponding phonemes are displayed in the top of the image. The alternating colors mark the onset and offset of each phoneme.

Appendix B

Résumé étendu

1 Introduction

La synthèse articulatoire a deux champs de recherche principaux, illustrés par la Figure B.1. Le problème direct, connu sous le nom de synthèse articulatoire de la parole, fait référence aux techniques utilisées pour synthétiser la parole à partir de formes articulatoires, et le problème inverse, connu sous le nom de l'inversion acoustique-articulatoire, qui fait référence à la reconstruction de la forme du conduit vocal à partir du signal acoustique. La synthèse articulatoire est une tâche complexe, mais elle permettra de produire un discours plus naturel et expressif, et peut bénéficier des systèmes de conversation à partir de la copie des mécanismes de production de la parole humaine en utilisant des techniques d'IA. Une étape importante de la synthèse articulatoire est la génération de formes réalistes du conduit vocal, un problème que nous dénommons "synthèse de l'articulation de la parole".

La description des mouvements du conduit vocal et des lieux d'articulation n'est pas une tâche simple. Même si la parole est un processus volontaire, elle se produit automatiquement. Typiquement, nous ne sommes pas conscients de notre articulation quand nous parlons. La synthèse des articulations apparaît donc comme une alternative efficace pour illustrer les articulations du conduit vocal pour les enfants, les patients recevant une aide d'orthophonie et les apprenants d'une deuxième langue.

La tâche présente nombre de difficultés. La production de chaque phonème dépend fortement du contexte, c'est-à-dire la coarticulation. La coarticulation fait référence à l'influence des phonèmes précédents (par inertie) et suivants (par anticipation) sur l'articulation d'un phonème central. En plus, l'articulation de chaque phonèmes doit respecter un ensemble de conditions physiques et acoustiques afin de produire le son correcte. L'articulation est aussi impactées par les phénomènes de compensation, quand le mouvement d'un articulateur compense la perte de mouvement d'un autre, e.g., comme c'est le cas de la langue et la mandibule affectées par le vieillissement.

Dans le contexte de l'apprentissage automatique, un grand défi qui affecte beaucoup la synthèse articulatoire est l'obtention d'une base de données fiable et suffisamment grande. La collecte de données articulatoires utilise typiquement l'imagerie médicale ce qui pose un certain nombre de difficultés. Les techniques d'imagerie médicale coûtent cher en comparaison avec l'acquisition des signaux de parole traditionnelle et posent plusieurs questions d'éthique afin de garantir la sécurité du sujet. En plus, il est nécessaire de traiter des images pour obtenir la géométrie correcte du conduit vocal.

La recherche articulatoire présente une grande opportunité pour améliorer la synthèse de la parole, faire

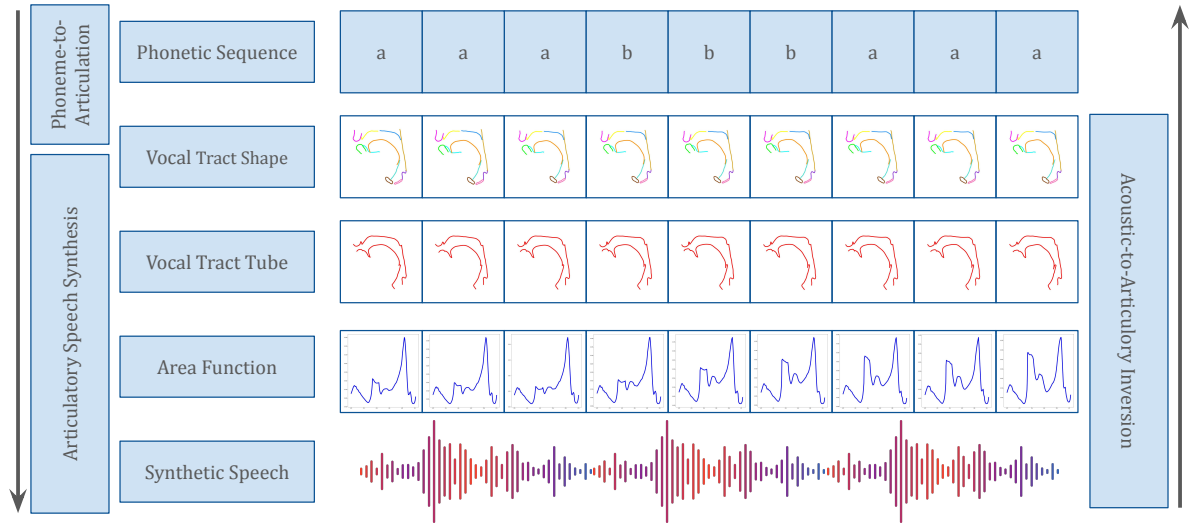


Figure B.1: Illustration de la synthèse articulatoire de la parole et de l'inversion acoustique-articulatoire.

progresser les systèmes de feedback visuel pour l'orthophonie et l'apprentissage des langues, et mieux comprendre la production de la parole. Cependant, elle pose aussi de grands défis que nous avons abordés pendant le développement de ce travail de thèse.

1.1 Motivation

Typiquement, la parole est un sujet auquel nous ne faisons pas attention. Quand nous parlons, nous ne sommes pas au courant de sa complexité. La différence temporelle entre le l'idée du message et la production de parole est une fraction de seconde. Cependant, la chaîne de la parole est élaborée. Parler nécessite l'orchestration des différents processus mentaux, psychologiques, physiques et sociaux. L' information est codée à de multiples niveaux quand elle passe d'informations neurocognitives dans le cerveau du locuteur à les ondes sonores et de nouveaux aux signaux électriques dans le cerveau de l'auditeur. Les aspects socio-comportementales, comme le ton, l'intention, les émotions, les caractéristiques démographique et l'état social, influencent profondément la perception des messages.

Mieux que la seule exploration acoustique de la parole, les méthodes multimodales permettent la compréhension de la structure et la fonction de l'appareil vocal. Les données articulatoires permettent de comprendre comment le conduit vocal transforme les signaux cérébraux en les ondes propagées dans l'air. Dans ce contexte, l'acquisition et le traitement de données articulatoires est un point essentiel pour les recherches sur la parole. De multiples techniques permettent d'observer la production de la parole, comme l'échographie pour le contour de la langue, l'articulographie électromagnétique (EMA) pour l'acquisition de la position de quelques capteurs, l'électropalatographie pour l'interaction entre la langue et le palais, et la cinéradiographie et l'IRM temps réel pour l'observation complète du conduit vocal.

Les données articulatoires sont essentielles pour déterminer la morphologie du conduit vocal et pour étudier et modéliser la communication orale afin de développer des applications dans de nombreux domaines. Hagedorn

et al. [12] a caractérisé les articulations du conduit vocal pour de la parole apraxique. L'apraxie affecte la bonne sélection et coordination temporelle des gestes vocaux; cependant, ces actions n'impactent pas la perception auditive. L'utilisation des articulations obtenues avec l'IRM temps réel facilite donc le diagnostic. En plus, Hagedorn et al. [13, 14] ont étudié la forme du conduit vocal et les stratégies compensatoires pour des patients ayant subi une glossectomie. La poursuite de ces recherches inclut l'utilisation de l'information articulatoire pour mettre en place la rééducation de la parole.

Par ailleurs, la recherche en articulatoire a un impact fort sur les recherches traditionnelles en traitement de la parole qui utilisent seulement le signal acoustique. Li et al. [15] a par exemple amélioré les systèmes de vérification du locuteur en ajoutant une information articulatoire (obtenue par inversion articulatoire) au signal acoustique. Srinivasan et al. [16] a exploré la robustesse des systèmes de reconnaissance automatique de la parole pour la parole normale et murmurée et il montré que l'information articulatoire est bénéfique dans les deux scénarios.

Ces travaux montrent que plusieurs de domaines tirent un avantage de la caractérisation ou de la synthèse de la forme du conduit vocal pendant la production de la parole. Les avancées concernant modèles articulatoires ont donc le potentiel de bénéficier à notre société à travers plusieurs champs de recherche.

1.2 Contributions Principales

À partir des considération générales sur la synthèse articulatoire et des motivations de cette recherche, cette thèse explore et développe les modèles d'apprentissage profond pour la synthèse articulatoire de la parole. Spécifiquement, notre contribution se focalise sur trois sujets principaux: le traitement des données articulatoires, la modélisation articulatoire et l'évaluation des modèles.

Traitement des Données Articulatoires

L'acquisition et le traitement des données articulatoires sont essentiels pour l'apprentissage automatique. Notre première source de données est l'imagerie par résonance magnétique (IRM) temps réel et il faut souligner que les images seules sont insuffisantes pour réaliser la synthèse articulatoire de la parole. Notre premier défi était de traiter cette images pour obtenir les contours de chaque articulateur du conduit vocal pour toutes les images de notre base de données. La littérature fournit plusieurs approches pour ce problème mais elles présentent souvent des faiblesses pour notre travail. Notre premier objectif de recherche a donc été *la conception d'un système fiable pour le suivi des articulateurs du conduit vocal dans des images IRM temps-réel.*

Modélisation Articulatoire

L'objectif central de cette thèse est *la synthèse de la forme du conduit pour une séquence de phonèmes à articuler.* La conception du système est difficile à cause de la dynamique complexe du conduit vocal humain, la grande variabilité des articulateurs, et les contraintes physiques et acoustiques à considérer.

Évaluation des Modèles

L'évaluation des modèles est cruciale pour l'apprentissage automatique. Il est important de mesurer les dimensions appropriées du problème et l'utilisation de mauvaises mesures conduirait à des modèles biaisés ou de mauvaise qualité. Les modèles d'apprentissage automatique optimisent l' objectif qu'on leur donne, et les réseaux de

neurones profonds cherchent donc le meilleur chemin pour optimiser leur objectif. La mauvaise définition de l'objectif d'apprentissage conduit à des conséquences souvent difficiles à anticiper et résoudre.

Comme notre recherche vise à prédire la forme du conduit vocal pour une séquence de phonèmes donnée, il est naturel de s'attendre que les modèles capturent le maximum d'information phonétique contenu dans la base de données. Finalement, le dernier objectif de recherche est de *quantifier l'information phonétique retenue par les contours des articulateurs et celle reproduite par le synthétiseur de la forme du conduit vocal.*

2 Segmentation des Articulateurs du Conduit Vocal dans l'IRM temps Réel

La caractérisation de la forme du conduit vocal est essentielle pour plusieurs domaines de la recherche de la parole et l'IRM temps réel est la modalité de donnée préférée pour de nombreux chercheurs. Cependant, pour la synthèse articulatoire, la géométrie exacte de la colonne d'air déterminée par les contours des articulateurs de la glotte aux lèvres est typiquement nécessaire.

La difficulté que nous avons rencontrée dans la littérature est que peu d'articles fournissent les contours de tous les articulateurs non-rigides et les articles montrent pas comment aborder la généralisation d'un système à plusieurs locuteurs. En plus, aucun article ne fournit une base de code publique pour l'utilisation et validation des ses méthodes. Cette thèse propose la segmentation des contours des articulateurs en utilisant un réseau de neurones profond suivi par des algorithmes de post-traitement afin d'obtenir une courbe qui décrit la géométrie de chaque articulateur individuellement.

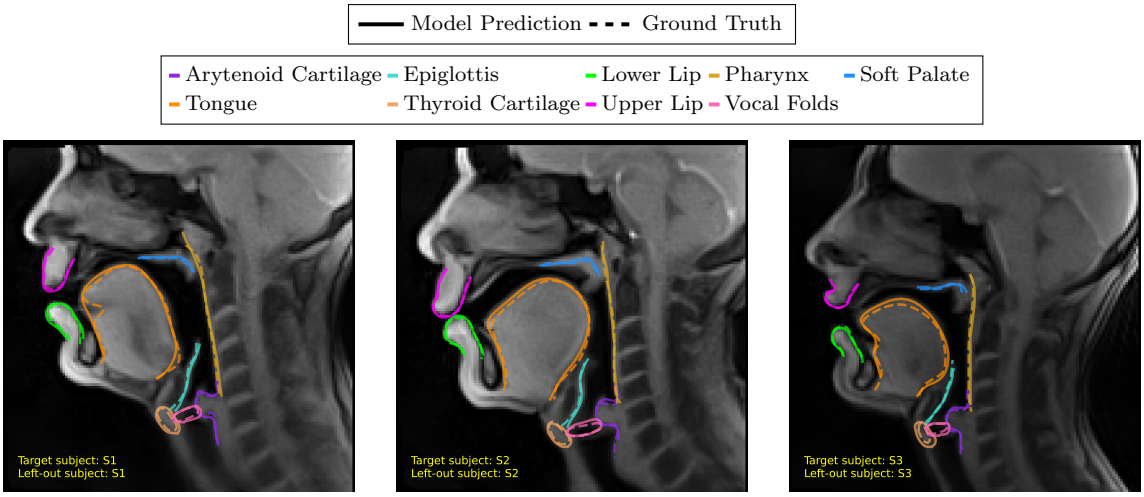
Les contributions principales de cette travail sont:

- La prise en compte des articulateurs non-rigides impliqués dans la production de la parole;
- L'évaluation de la généralisation à plusieurs locuteurs en utilisant un protocole de validation croisée "leave-one-out" (LOOCV);
- Le traitement d'une grande base de données d'IRM temps réel avec une erreur petite par comparaison avec un annotateur humain;
- La mise à disposition publique du système de segmentation qui permet l'utilisation et la validation de notre travail par la communauté scientifique.

Au final, nous avons proposé et évalué une méthode de suivi des articulateurs du conduit vocal à partir d'IRM temps réel (Figure B.2). Le système en deux étapes que nous avons proposé est capable de segmenter les contours des articulateurs non rigides avec une erreur maximale de 2.2 mm dans le protocole de validation croisée leave-one-out, qui teste la généralisation à de nouveaux locuteurs. Par ailleurs, pour le locuteur qui présente une différence anatomique importante et/ou une position de la tête éloignée des positions des autres sujets pour l'acquisition IRM, le modèle a pu être adapté avec seulement 10 images étiquetées.

3 Synthèse de l'évolution temporelle de la forme du Conduit Vocal

La synthèse articulatoire de la parole donne lieu à de nombreux défis, particulièrement le problème de la non-unicité, la normalisation de locuteur et la position des articulations critiques. La non-unicité signifie que plusieurs



(a) (b) (c)

Figure B.2: Échantillons d'IRM de trois sujet superposés aux contours prédits et de vérité terrain après régularisation b-spline. Le texte dans les images indique l'ID du sujet dans l'image (sujet cible) et l'ID du sujet laissé de côté lors de la formation du modèle qui a produit cette sortie. Cette figure montre comment les contours prédits se comparent aux contours de la vérité terrain pour chaque sujet.

configurations du conduit vocal produisent les mêmes caractéristiques spectrales. La normalisation du locuteur concerne les différences anatomiques qui nécessitent donc une normalisation afin de générer les formes pertinentes pour n'importe quel locuteur. Le défi est de séparer la variabilité liée aux stratégies de production de la parole et de celle liée à l'anatomie du conduit vocal. Les articulations critiques que nous avons utilisées viennent de la phonologie articulatoire, qui définit les gestes du locuteur et les scores de gestes pour articuler chaque phonème. Un articulateur est critique pour un phonème si le lieux d'articulation est nécessaire pour obtenir les bons traits phonétiques. Si un articulateur n'est pas critique, il est libre pour ce phonème. L'importance des ces défis est qu'ils conditionnent la synthèse des formes du conduit vocal pertinentes phonétiquement.

Cette thèse décrit notre approche pour la synthèse articulatoire de la parole à partir d'une séquence de phonèmes à articuler. Nous proposons deux modèles: uns approche "model-free", qui ne dépend pas d'un modèle articulatoire explicite pour la synthèse, et une approche basée sur un autoencodeur ("autoencoder-based"), qui utilise un autoencodeur comme modèle articulatoire du conduit vocal. Nous avons proposé comme système de base pour nos évaluation, un système qui utilise la position moyenne des articulateurs pour chaque phonème, sans prendre en compte les sons voisins.

Les contributions principales de cette travail sont:

- Le développement d'un modèle de réseau de neurones profond pour la synthèse articulatoire de la parole que ne dépend pas d'un modèle articulatoire du conduit vocal;
- Le développement d'un modèle articulatoire du conduit vocal basé sur un autoencodeur et l'utilisation de cet autoencodeur pour la prédiction de la forme du conduit vocal par un réseau de neurones;
- La comparaison des systèmes de synthèse articulatoire et la définition d'un nouvel état d'art pour la synthèse articulatoire.

Au final, nous avons présenté un système de génération de la forme du conduit vocal pendant la production de la parole. Les résultats mesurés à l'aide de la distance `P2CP` et par les variables du conduit vocal indiquent que

le système model-free et l'autoencodeur produisent des dynamiques similaires. Nous avons observé que le modèle autoencodeur est plus cohérent avec la notion d'articulation critique, une fois que la fonction de perte que nous avons créée permet l'introduction des informations numériques permettant d'évaluer des articulateurs critiques. Cependant, nous avons vu que le système model-free est plus facile à entraîner, et produit une dynamique plus stable temporellement – comme cela peut être observée dans les vidéos fournies comme matériel supplémentaire.

4 Évaluation des Modèles de Synthèse Articulatoire

L'évaluation des modèles de synthèse articulatoire est un défi à cause de la variabilité anatomique interlocuteur et de la variabilité des stratégies articulatoires. Deux métriques principales existent dans la littérature : les métriques qui utilisent la distance géométrique entre courbes et les métriques basées sur les variables du conduit vocal. La distance entre des courbes géométriques est facile à implémenter et interpréter, cependant elle devient difficile à utiliser pour plusieurs locuteurs sans effectuer avant la normalisation anatomique du locuteur. Les métriques qui utilisent les variables issues de la phonologie articulatoire du conduit vocal fonctionnent bien pour les consonnes, mais elles ne sont pas adaptées pour les voyelles.

Nous nous sommes inspirés de la littérature qui utilise la reconnaissance phonétique à partir des IRM et de données d'EMA pour évaluer la synthèse des formes articulatoires du conduit vocal à partir de représentations phonétiques. Nous avons donc entraîné et testé un reconnaisseur de phonèmes avec le signal acoustique de la parole afin d'obtenir un résultat de base pour la reconnaissance, qui mesure donc l'information phonétique totale contenu dans le signal. Par ailleurs, nous avons entraîné et testé le reconnaisseur de phonèmes avec les informations articulatoires obtenues à partir de l'IRM, qui mesure l'information phonétique donnée par les contours des articulateurs dans le plan médio-sagittal. Finalement, nous avons testé le reconnaisseur entraîné avec les articulations d'IRM sur les formes articulations produites par les modèles de synthèse que nous avons développés. La dernière expérience mesure l'information phonétique capturée et reproduite par notre synthétiseur de la forme temporelle du conduit vocal.

Les contributions principales de ce travail sont:

- La proposition d'une méthode simple et efficace pour évaluer les modèles de synthèse articulatoire;
- L'estimation de l'information phonétique contenue dans les données articulatoires obtenues à partir d'IRM temps réel et celle synthétisée par nos modèles articulatoires.

Au final, nous avons testé cette stratégie et observé que le signal articulatoire obtenu à partir des IRM donne des erreurs de reconnaissance de phonèmes (PER) comparables à celles obtenues à partir du signal acoustique et que la prise en compte de l'information de voisement améliore le taux de reconnaisance à partir des données IRM (à un niveau un eu plus bas que l'acoustique). Nous avons par ailleurs observé que la synthèse articulatoire "model-free" produit des articulations qui donnent un PER plus bas, donc meilleur, que le signal articulatoire original. Le résultat est surprenant, mais nous avons conclu que le modèle de synthèse articulatoire filtre le bruit et l'erreur de suivi qui existe dans le signal d'IRM, résultant ce qui conduit à des formes articulatoires temporelles plus faciles à reconnaître.

Finalement, la comparaison entre le système model-free et le système autoencodeur a montré qu'il existe une grosse différence entre les deux modèles. Comme nous l'avons noté, le système model-free produit des articulations plus stables temporellement, et cet effet est observé dans le PER de chaque système. Cette différence a montré que la reconnaissance des phonèmes est une méthode efficace pour mieux comprendre la synthèse des articulations et que cela permet d'évaluer les dimensions qui échappent aux métriques traditionnelles.

5 Conclusion

La synthèse de l'évolution temporelle de la forme articulatoire du conduit vocal est un problème complexe qui correspond à plusieurs défis, parmi lesquels les difficultés d'acquisition des données articulatoire, la modélisation du conduit vocal respectant la physique de production de la parole et les connaissances phonétiques, et enfin l'évaluation des modèles construits. La littérature couvre typiquement le problème direct, c'est-à-dire la synthèse articulatoire de la parole, ou le problème inverse appelé inversion acoustique-articulatoire.

Cette thèse a proposé une approche visant à générer les formes du conduit vocal pour une suite de phonèmes à articuler. Cette transformation a beaucoup d'applications pour la production de la parole, comme l' orthophonie et l'apprentissage d'une langue étrangère. En plus, la capacité de recréer les articulations de la parole qui prend en compte la coarticulation améliore le lien entre la synthèse de la parole et les processus acoustiques, conduisant ainsi à un discours plus naturel.

Les contributions de cette thèse concernent le traitement des données articulatoires, la synthèse de la forme du conduit vocal et l'évaluation des modèles de synthèse. Pour le travail futur, nous souhaitons améliorer le suivi des articulateurs dans les images IRM, en particulier pour les régions de contact entre articulateurs mobiles et fixes comme les incisives supérieures.

Concernant la synthèse d'articulations, nous souhaitons ajouter des contraintes physiques dans les modèles afin de garantir les lieux d'articulation réalistes pour chacun des articulateurs. En plus, un champ de recherche fructueux concernera sans doute l'exploration des "transformeurs" pour la synthèse des articulations afin de comprendre le lien et les influences entre les phonèmes et chaque partie du conduit vocal.

Finalement, les métriques d'évaluation reposant sur la reconnaissance automatique des phonèmes ouvre des nouvelles opportunités pour l'entraînement des modèles d'apprentissage profond de la synthèse articulatoire. Un fois que le reconnaisseur est entraîné pour plusieurs de locuteurs, il deviendra possible de créer une fonction de perte qui réalise implicitement la normalisation du locuteur, en rendant donc possible un apprentissage multi-locuteur. Cependant, il reste important de bien comprendre quelles sont les caractéristiques que le réseau de neurones utilise pour la reconnaissance de phonèmes et si ces caractéristiques reflète nos connaissances phonétiques.