On temporal constraints for deep neural voice alignment

Yann Teytaut

Doctoral Thesis from Sorbonne Université

Spécialité **Informatique**
École Doctorale Informatique, Télécommunications et Électronique (Paris)
Sciences et Technologies de la Musique et du Son (STMS UMR 9912)
Institut de Recherche et Coordination Acoustique/Musique (IRCAM)
Équipe Analyse et Synthèse des Sons

Subject of the thesis:

On Temporal Constraints for Deep Neural Voice Alignment

Author:
Yann Teytaut

Supervisor:
Dr. Axel Roebel

Reviewers
- Pr. Gaël Richard — Professor, Télécom Paris
- Dr. Emmanouil Benetos — Reader, Queen Mary University of London (QMUL)

Examiners
- Pr. Jean-Pierre Briot — Research Director, LIP6 (CNRS/SU)
- Dr. Emmanuel Vincent — ***Jury President*** — Research Director, Inria Nancy - Grand Est
- Dr. Rachel Bittner — Research Manager, Spotify Inc.
- Dr. Romain Hennequin — Head of Research, Deezer
- Dr. Chitralekha Gupta — Research Fellow, National University of Singapore (NUS)

À Jeannette Dat
née Berthe-Dominiquette Ladrix
.
Celle qui a tant donné
que des mots ne sauraient suffisamment l'exprimer.

To Jeannette Dat
born Berthe-Dominiquette Ladrix
.
The one who has given so much
that no words can express it enough.

"*Es pas l'òme que ganha es lo* ***temps***."
.
"*Ce n'est pas l'homme qui gagne, c'est le* ***temps***."
.
"*It is not the human who wins, it is the* ***time***."

Proverbe Occitan – Occitan Saying

Preface

The research presented in this thesis results from a long history of passion for sciences and music, and a strong desire to work in their interdisciplinary field. In high school, I already knew that I wanted to enroll in the ATIAM[1] (*Acoustique, Traitement du signal et Informatique Appliqués à la Musique*, *i.e.*, sciences applied to music) Master's program coordinated and housed at IRCAM[2], the Institute for Research and Coordination Acoustics/Music, which I did upon graduation from the IOGS[3] (*Institut d'Optique Graduate School*).

This is during the course of the ATIAM mandatory internship that I first discovered and got familiar with general music processing and the specific audio alignment task as Arshia Cont and Philippe Cuvillier gave me the opportunity to work on their score following system, *Antescofo*[4]. Ever since, the open challenges and the numerous and various applications offered by the alignment task have kept captivating me.

As I was willing to experience a journey within the academic field, I applied for the voice-related PhD position proposed by Axel Roebel from IRCAM's Analysis/Synthesis team, in the context of a new ANR (*Agence Nationale de la Recherche*) project. Entitled **ARS**[5], this project deals with the **A**nalysis and t**R**ansformation of singing **S**tyle and, to this aim, the development of alignment algorithms dedicated to voice − and not music as I used to at Antescofo − were necessary.

This is the story the present document covers. This journey began on January 20th, 2020 and shall end on June, 30th 2023. The thesis has been defended on July, 7th 2023 at Ircam[6].

This work has been fully funded by the French National Research Agency (ANR) project **ARS (ANR-19-CE38-0001-01)** and has also benefited from IRCAM's **UPI "ISiS Voices"**.

[1] https://www.atiam.ircam.fr/
[2] https://www.ircam.fr/
[3] https://www.institutoptique.fr/
[4] https://www.metronautapp.com/fr
[5] https://ars.ircam.fr/
[6] https://www.youtube.com/watch?v=O5RWUl_vZ9M&ab_channel=Ircam

Abstract

To listen, to respond, to make coincide, to coordinate, to adjust, to follow, to adapt, to be in unison, to synchronize, to align... The rich vocabulary dedicated to the correspondence of human activities shows the importance of their temporal organization. Human communication, multi-modal by nature, is fully concerned by this problematic since there exists a semantic gap between oral locutions and their symbolic sequences: how to interpret a written message without the vocal intonation? what performative style beyond a fixed musical score? This thesis proposes to uncover the complex underlying relationships between the audio and symbolic domains in order to reduce this gap through the fine study of the inherent temporality contained in voice recordings. The voice alignment task lies at the core of this objective, as it aims to determine the temporal occurrence of symbols that are assumed to be present in a voice signal. This work notably focuses on the development of an acoustic model, ADAGIO, capable of estimating such time-symbol links. Recent progress in deep learning have led to implement ADAGIO as a deep neural network in a powerful generic formalism: the "Connectionist Temporal Classification" (CTC). However, the great flexibility offered by CTC is undermined by its intrinsic lack of guarantees for temporally accurate predictions. Therefore, the key contributions of this research consist in reinforcing CTC with additional temporal constraints to improve the quality of the inferred alignments. To do so, three ancillary tasks of (1) spectral content reconstruction; (2) audio structure propagation; and (3) guided monotony are introduced and induce a positive impact on the alignment between voices, texts, and notes. Then, ADAGIO contributes to many practical applications via collaborations such as concatenative speech synthesis or the study of expressive production strategies at play for both social attitudes in speech and singing style in musical performances.

Résumé

S'écouter, se répondre, faire se coïncider, se coordonner, s'accorder, se suivre, s'adapter, être à l'unisson, se synchroniser, s'aligner... Le riche vocabulaire dédié à la mise en correspondance dans le temps des activités humaines montre l'importance que revêt leur organisation temporelle. La communication humaine, multi-modale par nature, est pleinement concernée par cette problématique puisqu'il existe un écart sémantique entre les locutions orales et leurs séquences symboliques : comment bien interpréter un message écrit sans l'intonation vocale ? quel style performatif au delà d'une partition musicale figée ? Cette thèse se propose de révéler et expliquer les complexes relations entre les domaines audio et symbolique afin de réduire cet écart grâce à l'étude fine de l'inhérente temporalité contenue dans les enregistrements vocaux. Au coeur de cet objectif, se trouve la tâche d'alignement de voix qui vise à déterminer l'occurrence temporelle de symboles supposés présents dans un signal vocal. Ces travaux s'intéressent tout particulièrement au développement d'un modèle acoustique, ADAGIO, capable d'estimer de tels liens temps-symboles. Les récents progrès en apprentissage profond amènent à implémenter ADAGIO sous la forme d'un réseau de neurones profond dans un puissant formalisme générique : la "Classification Temporelle Connectioniste" (CTC). Cependant, la grande flexibilité offerte par la CTC est mise en défaut par son absence intrinsèque de garanties de prédictions temporellement précises. Les contributions clefs de cette recherche visent à renforcer la CTC par des contraintes temporelles supplémentaires pour améliorer la qualité des alignements déduits. Pour cela, trois tâches annexes de (1) reconstruction du contenu spectral, (2) propagation de la structure audio, et (3) monotonie guidée sont introduites et induisent un impact positif sur l'alignement entre voix, textes, et notes. Dès lors, ADAGIO contribue à de nombreuses applications pratiques au travers de collaborations telles que la synthèse vocale concaténative ou l'étude des stratégies de production expressives en jeu tant pour les attitudes sociales dans la parole que pour le style de chant dans des performances musicales.

Acknowledgments

One cannot write the story of a thesis alone. It is written in the ink of progressive learning, scientific achievements, and the human relationships that made it possible and that developed during its course. In this respect, it would be impossible to begin this document without thanking all the people who have counted for and contributed to this story.

My first thoughts go to my family, whose love and benevolence have never failed me, and have been the foundation of my personal and professional development. Thank you to my mother and father, Annie and Jean-Claude, to all my brothers and sisters – Loïc, Hélène, Marie-Émilie, Julie and Laura – and to Brigitte, Didier and Françoise. I owe you everything.

Secondly, I'd like to express my deepest gratitude to my thesis supervisor, Axel Roebel, who placed his trust in me for this project, was always available, offered invaluable advice and provided inspiring understanding. I'd also like to thank Guillaume Doras, my "shadow" supervisor, who taught me so much while working with him. I'll never forget that. Thanks also to Arshia Cont and Philippe Cuvillier for this year of research in their company, where I knew I had found my field. It was a pleasure to do this thesis in the Sound Analysis-Synthesis team and to meet so many passionate and exciting people at Ircam during these years of thesis or Master ATIAM.

Thanks to Clément, my research brother, for all those moments, discussions, coffees, escapes and laughter. It was an honor, and a real inspiration, to live this adventure with you. Thank you to our entire *doctoral team* for all the memories, all too many BMs, "All night– all night long", top chef debriefs, Tata Burgers, cultural and not-so-cultural outings and mutual support. In particular, Constance for that ATIAM night writing, all those laughs and moments of non-verbal communication that have given me so much joy over the years, Victor for his patience, his imitations, his calls and voicemails and his scandals that I wouldn't change for anything in the world, Nadia for her natural and luminous goodness, the Korea-Japan visios, and the many pieces of chocolate I ate from her, Claire for her words and her sensitivity, the "cups of water" and her tent that stayed in the office for months, Baptiste for the maroilles, the "easy" piano solos and his "bip bap boup" experiments, Paul for his infectious dynamism and the start of

my cinema career, Valérian for one of the rarest sincerity and kindness I have ever met and for being the best ATIAM neighbor, Vincent for being an amazing cultural calender and opera companion, as evidenced by these numerous concerts together, and finally Pablo for his many personal and professional advice.

I have a special place in my heart for all my office colleagues: Gabriel, Hadrien, Lenny, Alice and Mathilde. My timing couldn't have been better. I'd like to thank all of you, in your own way, for enabling me to move forward. A special thank you to Lenny, the quiet force, for our discussions and his commitment. I am looking forward to our next Pizza. Thanks to Caio, Théo and Lucas for doing their internships/graduate projects with me. Thanks to the students I taught. Thanks to all the Ircam staff, whose kindness I will always remember: Deborah, Brigitte, Éric, Bruno, the PDS team for all their hospitality, and to all the people I can't mention here for lack of space but not of heart.

An immediate thought goes to all the people I met at/thanks to the Institut d'Optique Graduate School (IOGS) and today dear friends. Thanks to Fahim and Alice for your harmonies, to Perrine and Éric for Rennes and Haute-Savoie, to Bathilde, François and Yvan for always being there at major events, to Arianna and Lambert for our increasingly unexpected exchanges of stories, to François-Marie and Léna for simply being themselves and Jean-Baptiste for always being such a great friend. Thanks to all the BrBi's and VK's − I can't name you all, but my heart's in it. Thanks to Juliette for my fabulous first years in Paris and this road trip my thoughts often go back to.

Thank you Salomé. You have done so much that it would be impossible to simply sum it all up. I will never forget the year we turned 26.

I'm deeply touched to see that, despite the years, relationships can disrupt and remain strong. Thanks to Adeline for her constant support over the years and Disneyland, to Fanny for her sincerity and sweetness. Thanks to my prépa friends − the whole team, Élodie, Zakaria, Victor, Léa and Marie − for so many memorable moments that I often think back on. "Teytaut" thanks you sincerely. Thank you to the friendships from primary school, junior high and high school that have remained intact − Sophie for all those stories to retrace, from Breaking Benjamin to "My Liege" and visits during lockdowns, Estelle for her infectious joy, those "very chill" moments and our bursts of laughter, Caroline for being the best partner in S, the TPE, the INSA stays and so much more, and Anna, very humbly, for the music and the piano.

These acknowledgements end with the person to whom this thesis is dedicated: Berthe-Dominiquette or Jeanette. Not all heroines are household names. I am eternally grateful for everything she has given me and passed on to me. The memory that comes to mind is reciting the multiplication tables with her, which undoubtedly enabled me to take my first step towards scientific research.

Remerciements

L'histoire d'une thèse ne s'écrit pas seul. Elle s'écrit à l'encre des apprentissages progressifs, des acquis scientifiques, et des relations humaines qui l'ont permise et qui se sont nouées pendant son déroulement. À ce titre, il serait impossible de commencer ce document sans remercier toutes les personnes qui ont compté pour et contribué à cette histoire.

Mes premières pensées vont à ma famille dont l'amour et la bienveillance, qui n'ont jamais fait défaut, ont été les fondements de mon épanouissement personnel et professionnel. Merci à mère et à mon père, Annie et Jean-Claude, à mes frères et mes soeurs de sang et de cœur – Loïc, Hélène, Marie-Émilie, Julie et Laura – ainsi qu'à Brigitte, Didier et Françoise. Je vous dois tout.

Ensuite, j'adresse ma plus grande reconnaissance à mon directeur de thèse, Axel Roebel, qui m'a fait confiance pour ce projet, a toujours été disponible, de conseils précieux, et d'une compréhension inspirante. Merci aussi à Guillaume Doras, l'encadrant de l'ombre, qui m'a énormément apporté et appris. Je ne l'oublierai pas. Merci également à Arshia Cont et Philippe Cuvillier pour cette année de recherche dans leur entreprise, où j'ai su que j'avais trouvé mon domaine. C'était un plaisir d'effectuer cette thèse dans l'équipe Analyse-Synthèse des Sons et de rencontrer tant de gens passionnés et passionnant à l'Ircam au cours de ces années de thèse ou de Master ATIAM.

Merci à Clément, mon frère de recherche, pour tous ces moments, discussions, cafés, évasions et rires. C'était un honneur, et une réelle inspiration, de vivre cette aventure avec toi. Merci à toute notre *team doctorale* pour les souvenirs gravés, les trop nombreux BM, les "All night− all night long", les debriefs top chef, les Tata Burgers, les sorties culturelles et moins culturelles et le soutien mutuel. Notamment, Constance pour cette nuit blanche rédactionnelle ATIAM, tous ces rires et moments de communication non-verbale qui m'ont apporté tant de joie au fil des années, Victor pour sa patience, ses imitations, ses appels et messages vocaux et ses scandales que pour rien au monde je ne changerais, Nadia pour sa bonté naturelle et lumineuse, les visios Corée-Japon, et les nombreux morceaux de chocolats que je lui ai mangés, Claire pour ses mots et sa sensibilité, les "cups of water" et sa tente restée des mois dans le bureau, Baptiste pour

les maroilles, les solos de piano jugés faciles et ses “bip bap boup”, Paul pour son dynamisme communicatif et le début de ma carrière cinéma, Valérian pour l’une des plus rares sincérité et gentillesse que j’ai pu rencontrer et pour avoir été le meilleur voisin ATIAM, Vincent pour avoir été le meilleur agenda et compagnon culturel comme en attestent nos nombreux opéra, et enfin Pablo pour ses nombreux conseils personnels et professionnels.

Je pense et garde une émotion toute particulière pour tous mes collègues de bureau : Gabriel, Hadrien, Lenny, Alice et Mathilde. Je n’aurais pas pu mieux tomber. Je vous remercie, toutes et tous à votre façon, d’avoir permis mon avancée. Un merci tout particulier à Lenny, la force tranquille, pour nos discussions et son investissement. J’ai hâte de notre prochaine Pizza. Merci à Caio, Théo et Lucas d’avoir effectué leurs stages/projets de fin d’étude avec moi. Merci aux étudiants à qui j’ai pu enseigner. Merci, enfin, à tout le personnel Ircam dont la gentillesse ne sera pas oubliée : Deborah, Brigitte, Éric, Bruno et toute l’équipe PDS pour leur hospitalité, et toutes les personnes manquantes à cette liste faute de place mais pas de cœur.

Une pensée immédiate va à toutes les personnes rencontrées à ou grâce à l’Institut d’Optique Graduate School (IOGS) et aujourd’hui amis chers. Merci à Fahim à Alice pour vos harmonies, à Perrine et Éric pour Rennes et la Haute-Savoie, à Bathilde, François et Yvan pour leur présence sans faille aux grands évènements, à Arianna et Lambert pour nos échanges d’histoires toujours plus inattendues, à François-Marie et Léna d’être purement et simplement eux-mêmes et à Jean-Baptiste pour toujours avoir été un merveilleux ami. Merci à toutes les BrBi et VK – je ne peux pas tou-tes vous citer mais le cœur y est. Merci à Juliette pour mes premières années parisiennes qui ont été fabuleuses et ce *road trip* auquel souvent je repense.

Merci à Salomé. Tu as tant fait qu’il serait impossible de le résumer. Jamais je n’oublierai nos 26 ans.

Je suis profondément touché de voir que, malgré les années, les relations peuvent perturder et rester fortes. Merci à Adeline pour la constance de son soutien au fil des années et pour Disneyland, à Fanny pour sa sincérité et douceur. Merci à mes ami-es de prépa – toute l’équipe Élodie, Zakaria, Victor, Léa et Marie – pour tant de moments marquants auxquels souvent je repense. « Teytaut » vous remercie sincèrement. Merci aux amitiés de primaire, collège et lycée qui sont restées entières – Sophie pour tous ces récits à retracer, de Breaking Benjamin à « My Liege » en passant par les « motifs familiaux impérieux », Estelle pour sa joie communicative, ces moments « très chills » et nos éclats de rire, Caroline pour avoir été la meilleure binôme de S, le TPE, les séjours INSA et tant encore, et Anna, très humblement, pour la musique et le piano.

Ces remerciements se concluent par la personne la personne à qui cette thèse est dédiée : Berthe-Dominiquette ou Jeanette. Toutes les héroïnes ne sont pas connues du grand public. Pour tout ce qu’elle a su m’apporter et me transmettre, ma reconnaissance est éternelle. Le souvenir qui me vient concerne la récitation avec elle des tables de multiplication, me permettant d’effectuer, sans doute, mon premier pas vers la recherche scientifique.

Table of Contents

Chapter 1

Introduction

*"If you knew **Time** as well as I do," said the Hatter, "you wouldn't talk about wasting it."*

The Hatter in Alice's Adventures in Wonderland
– Lewis Carroll

Time.

A name, a concept, a notion so common and yet so disconcerting.

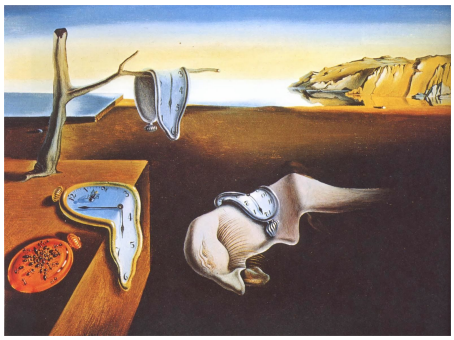
Not only has time obsessed artists, philosophers, or even the common, ordinary human sensitive to, concerned by, or simply aware of the finitude of the human condition, but it has also captivated scientists of various disciplines.

Representing time – whether objective ("physical") or subjective ("lived") – has especially stimulated many minds. It is a recurrent source of inspiration in arts, maybe as a necessity to express its omnipresence and the lack of means to control it. Because of its absence of any concrete *form*, time has been portrayed through common related objects (*e.g.*, hourglass or clocks) by painters or embodied in characters by writers, as shown in Figure 1.1, often dealing with the underlying anguish of time flying and of the fleeting nature of life and memory.

The question of time representations is also of highest importance for scientists eager to study the evolution of physical phenomena – as most physical properties are not static but rather dynamic – and for general *ordering* between events. Such objectives have led to define a reference for time *measurements*, the *second*, and incited to gather *knowledge* about time itself.

(a) *Vanité* or *Allégorie de la vie humaine* (Philippe de Champaigne, 1644).

(b) *The White Rabbit* (Sir John Tenniel, 1890) for *The nursery Alice* (Lewis Carroll, 1889-1890).



(c) *La persistencia de la memoria* (Salvador Dalí, 1931).

Figure 1.1: Time representations in arts.

The probably never-ending quest for the deep understanding of time has repeatedly revealed its special nature. Tracing back its history, questioning its origin in astrophysics, trying to model its evolution or report on its impacts on physical observations with mathematical laws – time was shown to be, and still remains, an intriguing concept full of secrets.

Indeed, among the four dimensions humans are familiar with – three spatial coordinates, usually denoted with the Cartesian (x, y, z) or Polar (r, θ, φ) coordinate systems, and time t –, time is the only dimension that cannot be "touched", crossed forward *and* backward, but rather that humans inexorably experience. The noticeable theories of Special Relativity and General Relativity, focusing on time-space relationships, demonstrate that time measurements are not even universal but are subject to the quite counter-intuitive *time dilation* effect such that time passes at different rates for two clocks (observers) in relative motion or submitted to different gravitational potentials – perhaps as if the very essence of time could not be universally determined but should rather maintain some level of inaccessibility.

This manuscript, far from considering time *per se* as a core topic of study, rather proposes to investigate some of its implications, which are ubiquitous in one's daily life. Many signals, as observed quantities in the surrounding world, are non-stationary such that their observations vary with the passing of time. Among the wide variety of existing signals with this property, this thesis clearly focuses on characterizing audio observations – taking various practical shapes between sounds, music, speech and singing, each with their intrinsic and fascinating diversity.

Notably, it is undeniable that time plays a crucial role in the way humans interact with one another – and particularly through audio signals. The very vocabulary of verbal communication but also of musical exchanges, is highly correlated with temporality: *listening to, responding, following, coordinating, adapting, synchronizing, aligning...* A deeper look at this temporal organization may thus reveal some of the characteristics at the heart of human communication.

The means of communication are, for humans, fundamentally *multi-modal* in the sense that the *same* idea can be expressed by *several media* and be represented in *different formats*. For example, a primarily way of expression is the *voice* which allows one to talk or sing by emitting sounds. But these locutions are very often associated with some underlying representation(s) of *symbolic* nature. Reciting a poem, reading a story out loud, giving a speech or a lecture, usually imply the existence of *written texts*. Similarly, performing a musical piece, acting in a musical or singing in an opera may be dependent on *music scores* and music theory.

Yet, albeit originally characterizing the same *message* to communicate, transitions between these diverse modalities are far from being straightforward. There are, indeed, for a single symbolic text, many ways of expressing it through the voice which can lead, for example, to humorous situations or misunderstandings when only a written message is available but the voice, intonation and *temporality* are missing. In the same way, a music score, although fixed, allows for multiple interpretations so that each concert and performance turns out to be unique.

The *semantic gap* precisely relates to these differences between two descriptions of a message or concept in different representations − *e.g.*, sounds, letters or symbols. In the very words of Andreas Hein (2010), the semantic gap corresponds to "*the difference in meaning between constructs formed within different representation systems*".

In the context of the **ARS** project − **A**nalysis and t**R**ansformation of **S**inging style, ANR-19-CE38-0001-03 (http://ars.ircam.fr) −, which has fully funded this thesis, one of the core objective was to develop means that allow describing the singing style in popular music. Given the multi-modal nature of music and especially recordings featuring singing voice, as mentioned, the idea to investigate algorithms capable of jointly and comparatively manipulating symbolic sound descriptions with arbitrary alphabets and audio recordings, saw the light of day.

By keeping these algorithms as generic and general as possible, they would not only allow studying singing performances but also, in a way broader scope, allow bridging the semantic gap − notably by exploiting the inherent temporality contained in voice signals.

1.1 Thesis research scope

Therefore, the scope of this thesis is to uncover the complex underlying temporal relationships between several representations conveying communication messages, with a focus on *voice* signals which, by nature, feature strong variations over time. The main research problem at the core of this challenge is *temporal voice alignment* or *synchronization*. It is studied as a goal for designing systems capable of revealing temporal information from voice data, and as a means for the analysis and synthesis of voice signals and their intrinsic expressivity.

Concretely, the *alignment* task aims to determine the precise time positions of a symbolic sequence, whose symbols are known to constitute the content of a given audio signal. The problem requires an audio file and a symbolic sequence as inputs. Note that the alignment is distinct from *recognition*, which consists in finding the symbols present in an audio signal.

Deep Neural Networks (DNN) are worthy candidates to address the alignment problem as they allow learning complex mapping functions between different domains (here, audio and symbolic) entirely informed by data. To fully benefit from this *data-driven* approach, no task-specific contextual information is provided besides the ones that can be directly inferred from the inputs at disposal (*i.e.*, voice signals and symbols).

A central objective is to develop an algorithm working for arbitrary symbol sets such as characters, phonemes, or even music notes, that does not need any specific information about the problem at stake. This way, the system can handle different languages, speakers, audio and voice qualities (speech and singing), and ideally arbitrary audio length. The main idea in this thesis is that the integration of general (*i.e.*, non-task specific) *temporal constraints* directly in the design and data modeling stage of DNNs must result in a better alignment accuracy, given the primary importance of time in the processes discussed.

Addressing such a research problem requires to raise and tackle practical and/or research questions. Three of them (and their corresponding sub-questions) are answered in this document.

Question 1 (Q1). *Temporal voice alignment – what?*

- *What* is the temporal alignment task in general and for voice?
- *What* kind of representations can be used to align voice data?

Question 2 (Q2). *Temporal voice alignment – how?*

- *How* to develop a system for the temporal alignment of voice?
- *How* to extract suitable temporal information from audio to reinforce alignment accuracy?

Question 3 (Q3). *Temporal voice alignment – why?*

- *Why* is temporal alignment of interest in various research communities?
- *Why* does temporal voice alignment lead to numerous research applications?

Q1 – *what?* is a practical question necessary to get familiar with the task addressed in this thesis. **Q2** – *how?* is the research question that aims at investigating ways to achieve temporal voice alignment. **Q3** – *why?* is another practical question thought to better understand the need for temporal alignment, all in all giving motivation based on examples demonstrating the potential of such algorithms.

1.2 Summary of contributions

Subsequent to these interrogations, this thesis has proposed solutions and led various studies for their exploration and answering. This section provides a summary of these contributions and applications.

*ADAGIO – A system for the **A**utomatic **D**eep **A**li**G**nment of v**OI**ce*

The core proposal of this work is the development of **ADAGIO** – a system dedicated to the **A**utomatic **D**eep **A**li**G**nment of v**OI**ce. It is an *acoustic model*, *i.e.*, a model capable to temporally estimate and represent the symbolic content associated with an audio recording. Built upon recent advances in deep learning for voice analysis research, ADAGIO is implemented as an end-to-end Deep Neural Network (DNN) in a Connectionist Temporal Classification (CTC) training scenario. Its generic nature allows aligning voice signals with symbolic sequences as long as they describe successive events occurring in the audio.

Temporal constraints for acoustic model training

The second main proposal of this thesis is the definition of additional temporal constraints integrated during the training phase of the acoustic model, with the aim to better capture the temporal relationships and lead to an overall higher alignment quality. These constraints are:

- **Spectral content reconstruction** with the claim that a precise temporal reconstruction of the relevant spectral information (*i.e.*, spectral *envelope* or *excitation*) is heavily dependent on the voice alignment;
- **Temporal structure preservation** with the claim that shared similarity patterns, informing on the local temporal structure, are expected to be found in the original voice signals and in the alignment predictions;
- **Guided time-symbol monotony** with the claim that a pertinent alignment must highlight a clear monotonic path between the voice data sequences.

Validation on voice-to-text and voice-to-notes alignments

As a third contribution, experiments are conducted to evaluate ADAGIO enhanced with the above-mentioned temporal constraints and validate its generic nature and potential. Voice-to-*text* alignment is tackled at various granularities. Word-level and syllable-level alignment are obtained by aligning *graphemes*, and phonetic alignment through the alignment of *phonemes*.

Two evaluation datasets, one for speech (20 mn audio with manual word-level annotations) and one for singing voice, are proposed to the community. The robustness of the model to music instrumentals and even imperfect transcripts is demonstrated through a series of quantitative evaluations. Then, the voice-to-*notes* alignment problem is also tackled and the feasibility of aligning voice music scores with audio (despite background music) with ADAGIO is confirmed. However, this note alignment task can still be improved at the time of writing.

Collaborative applications to voice analysis and synthesis

Last but not least, ADAGIO has been employed in many practical collaborations in diverse research contexts such as automatic speech analysis, singing voice synthesis and musicological studies. The different applicative contexts in which ADAGIO has already been used are listed below:

- **Concatenative singing voice synthesis** with the integration of new voices for ISiS (Ircam Singing Synthesizer) thanks to their phonetic alignment with ADAGIO;
- **Production strategies of social attitudes** with a core component, offered by the phonetic alignment of expressive speech with ADAGIO, dedicated to study the temporal aspects involved when speakers aim at conveying precise social intentions;
- **Musicological singing style analysis** with the development and future online release of a complete pipeline allowing musicologists to study expressivity in sung performances notably thanks to *both* syllables *and* notes alignments made possible by ADAGIO;
- **Very long audio alignment** for which ADAGIO, as an acoustic model, is coupled with a linear memory decoding module, to synchronize very long recordings (*i.e.*, several hours, *e.g.*, entire music playlists or audiobooks) with their wide text transcripts.

1.3 Dissertation outline

The Figure 1.2 gives a global overview over the organization of the document. It also shows how the different parts and chapters relate to the raised practical and research questions. Concretely, besides this introduction (Chapter 1) and an overall conclusion (Chapter 8), this work is divided into three parts, each composed of two chapters.

- Part I – Voice alignment: context & background

  In this part, all the notions necessary for a complete reading of the whole manuscript are introduced by getting familiar with the task of temporal alignment of vocal signals. In this perspective, the practical motivations of a research work focused on temporal voice alignment and an extensive literature review are presented. More precisely:

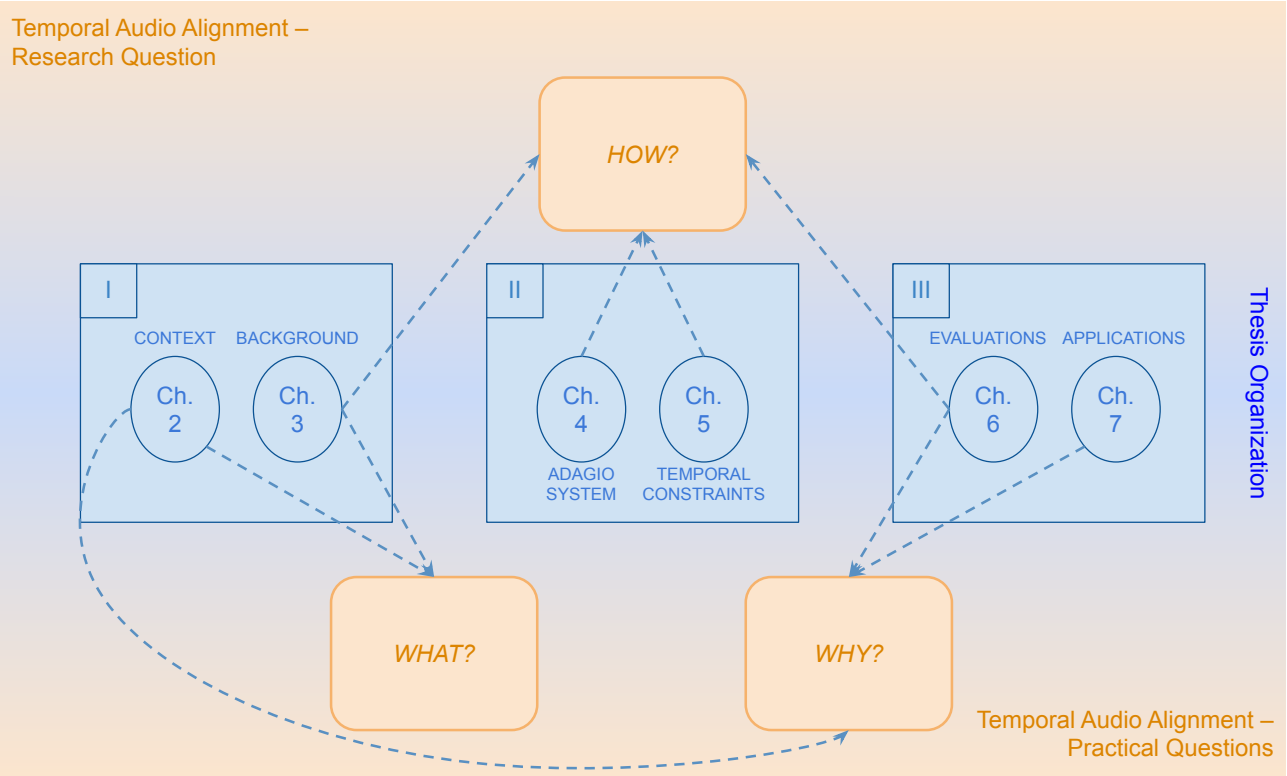


Figure 1.2: Outline of the dissertation and its relationships with the practical and research questions raised by temporal voice alignment.

— Chapter 2: Context and applicative motivations explores the notions related to human communication, and in particular vocal communication, to introduce audio time-frequency representations and symbolic sequences manipulated in this work (**Q1** – *what?*). Following their presentation, the task of their temporal alignment is defined (**Q1** – *what?*), as well as the opportunities offered by such synchronizations in the voice community (**Q3** – *why?*), thus making the link between the scientific disciplines mentioned and the technical tools to be presented in the next chapter.

— Chapter 3: Scientific background for voice alignment proposes a systematic review of the state of the art associated with voice alignment by presenting mathematical concepts and models from the literature. The specific task of voice alignment is formalized (**Q1** – *what?*) and then the various strategies used for its resolution are presented (**Q2** – *how?*). Recent approaches, including this thesis, are based on artificial intelligence, which motivates a summary of the key notions of deep learning. This work will concretely adopt a Connectionist Temporal Classification (CTC) training strategy whose specificities are detailed.

- Part II – Contributions: time-constrained neural voice alignment

  This part introduces a deep neural voice alignment algorithm, denoted ADAGIO, which, together with multiple losses imposing temporal constraints, is the main contribution of the thesis. This part, consequently, fully addresses the research question **Q2** – *how?* More precisely:

  — Chapter 4: ADAGIO, an acoustic model for temporal voice alignment exposes the acoustic model at the heart of this work – ADAGIO, a system dedicated to the "**A**utomatic **D**eep **A**li**G**nment of v**OI**ce". Its convolutional architecture and comparisons with other existing and previous systems are made.

  — Chapter 5: Temporal constraints for alignment enhancement aims at improving ADAGIO through the introduction of additional temporal information. Supplementary modeling objectives are proposed to reinforce the temporal coherency of the predictions and alignment quality. These contributions include a temporal audio reconstruction, propagation of the temporal structure, and temporal-sequential monotony assurance.

- Part III – Outcomes: experiments & applications

  This part, finally, focuses on the results of this thesis through the evaluations of the proposed models in comparison to relevant baselines, demonstrating the interest of the contributions, and by putting ADAGIO into practice. More precisely:

  — Chapter 6: Evaluations of deep voice alignment conducts a set of experiments evaluating the accuracy of the ADAGIO algorithm, an essential step for determining the relevance of any synchronizer (**Q2** – *how?*). This evaluation procedure is applied on several cases of practical interest (**Q3** – *why?*). Performance comparisons with relevant baseline algorithms demonstrate the robustness of ADAGIO and the proposals defended in this thesis.

  — Chapter 7: Applications and collaborations is dedicated to the practical and scientific contributions made possible by ADAGIO, by means of describing applications done in collaboration with other researchers (**Q3** – *why?*).

Chapter 8 features an overall conclusion to this work by summarizing the content of the manuscript and associated contributions as well as eventually answering to the three practical and research questions.

1.4 List of publications

The list of research papers that have been written during this PhD and accepted for publication in peer-reviewed national/international conferences and journal is presented below.

(Teytaut and Roebel, 2021)	**Y. Teytaut** and A. Roebel. Phoneme-to-audio alignment with recurrent neural networks for speaking and singing voice. In *Proceedings of Interspeech 2021*, pages 61–65. International Speech Communication Association; ISCA, 2021.
(Loiseau et al., 2022)†	R. Loiseau, B. Bouvier, **Y. Teytaut**, E. Vincent, M. Aubry, and L. Landrieu. A model you can hear: Audio identification with playable prototypes. In *23rd International Society for Music Information Retrieval Conference* (ISMIR 2022), 2022.
(Salais et al., 2022)	L. Salais, P. Arias, C. Le Moine, V. Rosi, **Y. Teytaut**, N. Obin, and A. Roebel. Production strategies of vocal attitudes. In *Interspeech 2022*, pages 4985–4989. ISCA, 2022.
(Teytaut et al., 2022)	**Y. Teytaut**, B. Bouvier, and A. Roebel. A study on constraining connectionist temporal classification for temporal audio alignment. In *Interspeech 2022*, pages 5015–5019. ISCA, 2022.
(Doras et al., 2023)	G. Doras, **Y. Teytaut**, and A. Roebel. A linear memory ctc-based algorithm for text-to-voice alignment of very long audio recordings. *Applied Sciences*, 13(3):1854, 2023.
(Teytaut et al., 2023)	**Y. Teytaut**, A. Petit, C. Chabot-Canet, and A. Roebel. A musicological pipeline for singing voice style analysis with neural voice processing and alignment. In *Journées d'Informatique Musicale* (JIM 2023), Saint-Denis, 2023.

Remark: the publication marked with †, as a collaborative and fruitful side project between vision and audio communities, is *not* concerned with temporal voice alignment − although it was interested in defining relevant *temporal* sound transformations to address speaker/*voice* identification. Its content, therefore, will *not* be discussed in this manuscript.

Part I

Voice alignment: context & background

Chapter 2

Context and applicative motivations

*"Music is nothing else but wild sounds civilized into **Time** and tune."*

– Thomas Fuller

This chapter presents the general context associated with this thesis, inscribing this research into larger and more specific fields of study, and the main practical motivations behind the development of temporal alignment algorithms for voice processing.

The starting point of this thesis is the *human communication* theory, and in particular vocal communication, as exposed in section 2.1. It allows introducing two major modalities of messages that human exchange when interacting, that are in line with this research: symbolic sequences and voice signals. Notably, representing the temporal evolution of the parameters contained in such signals is a mandatory step to further focus on temporal aspects. This requires *voice signal processing* background, given in section 2.2. Then, the two *transcription and alignment* tasks, at the heart of the manuscript as they link symbolic and audio representations, are deeply introduced in section 2.3. Finally, section 2.4 proposes an overview of the *practical applications* offered by the temporal synchronization between voice and symbolic information, motivating a research work in this domain. The chapter is summarized in section 2.5.

2.1 Human communication

Communication is an essential aspect of human life and has been crucial in the evolution of humans as a species. From hearing the voice of their mother before they are born to daily life implications − even up to summarizing years of research in PhD manuscripts −, humans communicate to connect and maintain relationships with one another, share information and knowledge. It is by communicating that humans bring about a common interpretation of the world in which they live, their environment, which makes possible their collaboration in the broadest sense. As a result, the study of human communication is of highest interest.

To this aim, a mathematical formalization of the concept of communication, introduced by Shannon (1948) in the form of the *communication theory*, is first presented. It allows defining the generic and fundamental notion of *message* necessary in all future investigations. A message can contain information of numerous natures, yet this thesis is particularly interested in *symbolic* and *oral* messages whose brief writing history and involved physical production mechanisms, respectively, are further exposed. In addition to this core part, humans are able to induce and interpret many variations when communicating a message − a short overview on this *expressivity* is proposed.

2.1.1 Communication fundamentals

Shannon*'s theory of communication*

A theoretical approach to understand communication can be traced back to 1948, when Claude Shannon introduced the concept of the communication chain − a model for comprehending the process of transmitting and receiving information, also known as the Shannon *model of communication* (Shannon, 1948).

According to this theory, the communication process involves six elements, as depicted in Figure 2.1. A sender emits a message m which is encoded by an encoder $\mathcal{E}$, transmitted over a communication channel and eventually decoded by a decoder $\mathcal{D}$ and intercepted by a receiver. If communication has been successfully achieved, the decoded message $(\mathcal{D} \circ \mathcal{E})(m)$ must be exactly the original message m.

However, this cannot be systematically guaranteed as there are sources of *noise* along the transmission chain, blurring information throughout the propagation of the message, and altering the effectiveness of the communication process. Noise can take many forms from physical (*e.g.*, interferences on a telephone line) to psychological (*e.g.*, misunderstandings or distractions preventing the receiver from fully comprehending the message).

The notion of *noise* is not static and highly depends on the message that is transmitted. For example, in different contexts, musical background can be considered as noise (*e.g.*, hearing someone talking in a concert) or as the key information to decode (*e.g.*, focusing on the construction of a music piece).

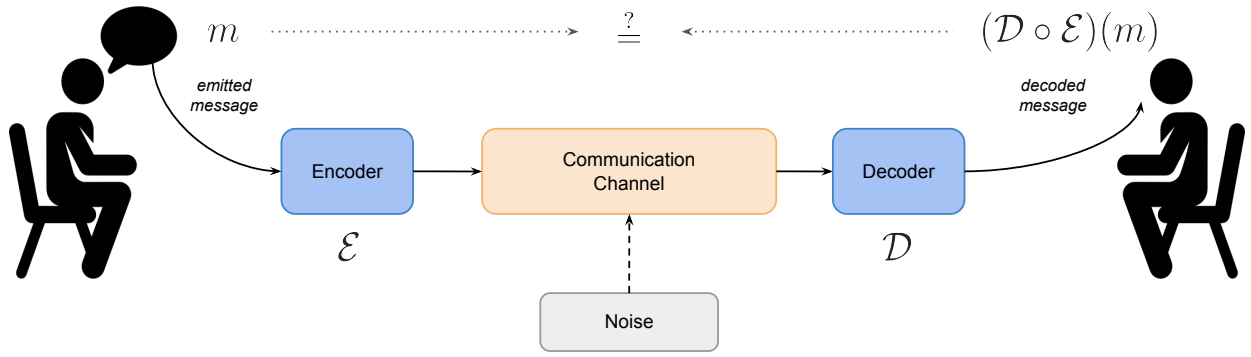


Figure 2.1: SHANNON's communication chain.

Communication message

The *message* m is a core element of the communication chain as it contains the information one wants to transmit. Figure 2.2 proposes a succinct overview of what messages are made of.

It can be summarized as follows:

- On the one hand, there is the central part of the message, its *content* and its *denotation*, which precisely relates to its strict meaning, *i.e.*, the sense and ideas that are to be communicated. This implies the existence of an underlying system – a *language* – defining a set of rules, admitted and shared by a group of people that allow to state understandable and meaningful messages. Human linguistics (COOK, 2003) and music theory (BESSON and SCHÖN, 2001) are examples of languages of interest in this work.

- On the other hand, there is always *meta*-information in a message, which is divided into two main categories: (1) the *para*-linguistic which *connotes* the strict meaning through, *e.g.*, expressivity and emotions; and (2) the *extra*-linguistic which places the speech in a context (*e.g.*, geo-socio origin, age, communication style). These extra- and para-linguistic aspects of communication will have their importance for applicative motivations – they are therefore highlighted in section 2.1.4.

Message modalities in voice research

Much like the information it contains, the nature of a message is highly variable. For a given language, indeed, a message can be expressed in diverse *modalities*.

In this thesis with a focus on the relationships between audio and texts, two well-known human communication modalities are ubiquitous: *oral* and *symbolic* modalities.

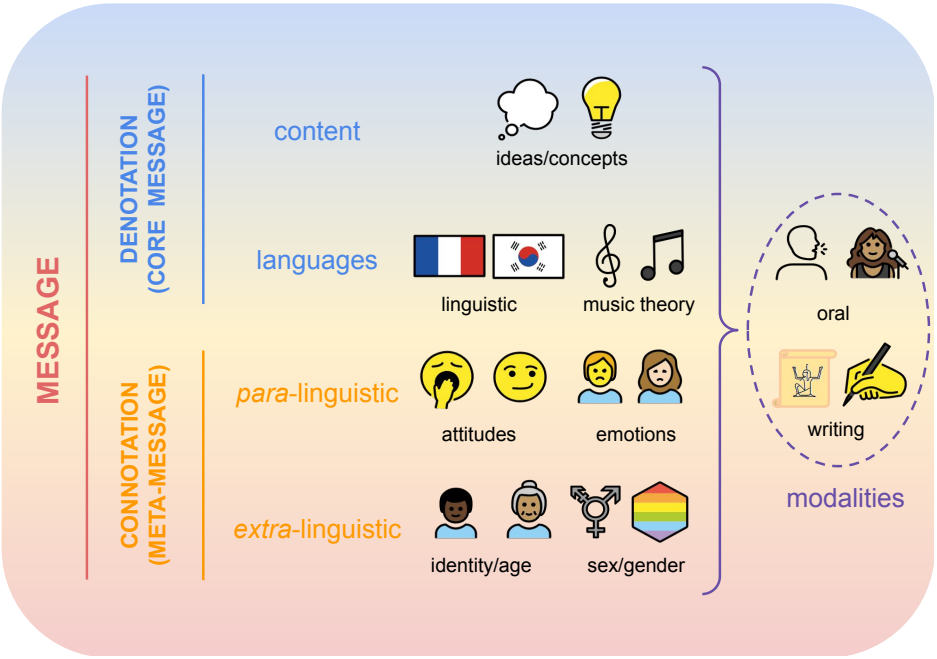


Figure 2.2: Content of communication messages. Illustrations from https://openmoji.org/

Symbolic messages are dependent on the existence of a *writing system* and are the focus of section 2.1.2. *Oral* messages, such as speech and singing, are produced by humans thanks to their voice via physical body mechanisms that are presented in section 2.1.3.

For the sake of completeness, even though beyond the focus of this thesis, it is worth mentioning that other types of modalities are studied in their associated literature – *e.g.*, facial motion or body gesture (Fares et al., 2021).

2.1.2 Writing and symbolic modality

Writing is a representation of a language through the use of visual *symbols* such as pictograms, figures, letters or characters. It is a system that allows people to communicate messages and transmit ideas in a *symbolic* form – through written documents. The ensemble of all symbols constituting a writing system is known as an *alphabet* $\mathcal{A}$.

A brief history of writing

The history of writing (Fischer, 2003) dates back to ancient civilizations, where it was a means of recording, preserving as well as transmitting information. Thanks to writing, humans were able to share notions beyond time and space, enabling the exchange of knowledge and the growth of civilizations.

Figure 2.3: Writing systems of some modern languages. The meaning is "***time***" for all words.

One of the earliest known forms of writing is *cuneiform*, which involved pressing a reed stylus into clay tablets to create symbols representing objects, words, and ideas. It was used by the Sumerians in ancient Mesopotamia (4000 BCE) to keep track of a variety of content (*e.g.*, religious texts, legal codes, etc.). In ancient Egypt (3000 BCE), writing was also an important tool of communication and record-keeping. The ancient Egyptians relied on *hieroglyphics* – a system of writing based on pictorial symbols representing words and ideas. Hieroglyphics were primarily inscribed on stone or other durable materials, such as tombs and temple walls.

With the passing of time, writing systems also evolved and adapted to the needs of different languages (Schmandt-Besserat, 2014). For example, the Phoenicians (2000 BCE) developed an alphabet leading to a simpler and more efficient system of writing: individual symbols were meant to represent *sounds* rather than whole *words* or *concepts*. The Phoenician alphabet served as the basis for the Greek alphabet and other upcoming writing systems.

Throughout history, writing played a crucial role in the development and spread of language and culture – and music notation fully inscribes itself in these considerations (Strayer, 2013). Writing, though, is not an inherent part of a language: most spoken languages, actually, have no standardized written forms or no written system at all. For illustration purposes, Figure 2.3 displays some contemporary writing systems.

Symbolic modality of communication

The research presented in this manuscript deals with languages that do have a well-defined *alphabet* $\mathcal{A}$ to write documents based on linguistic or grammar rules. This includes recent versions of unsigned living languages (*e.g.*, English, French, Greek, etc.) and classical musical notation and notes from music theory.

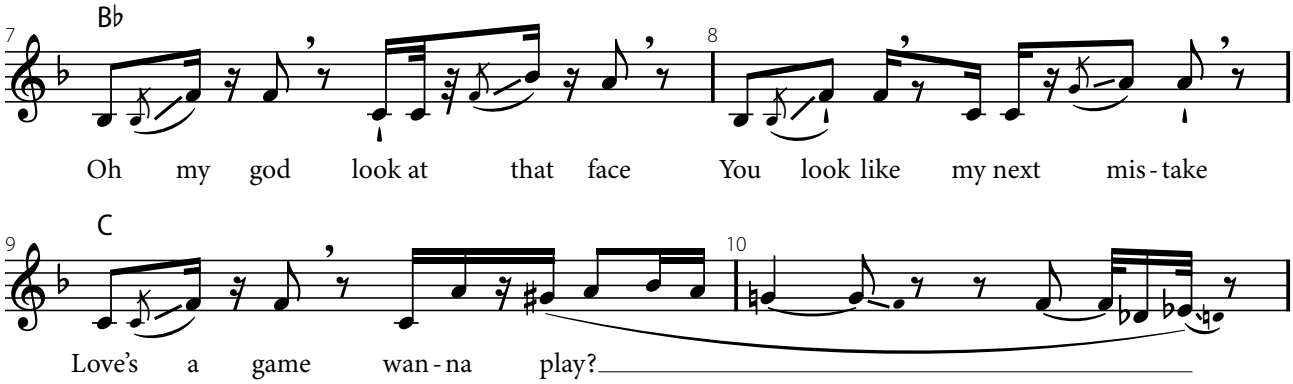


Figure 2.4: Singing music sheet features both text and note as symbolic information. Taylor SWIFT's "Blank Space", *1989* (2014), measures 7−10 (15:30−25:30). Transcription done by Antoine PETIT. Courtesy of Antoine PETIT (PETIT, 2022).

In a *textual* context − A *sequence* (or *text*) is defined as an ensemble of one or more of written message(s). Characters from the alphabet $\mathcal{A}$, which are referred to as *graphemes*, allow to constitute *syllables* or *words*, that compose *sentences*, leading to *paragraphs*. The *Latin alphabet* $\mathcal{A}_\ell$ will be primarily used as common to many spoken languagues. It contains all the basic latin characters/graphemes (a, b, c, etc.) and is augmented with the digits (0, 1, 2, etc.) as well as a *space* (ø) to separate words. All in all, it is defined as a set of 37 symbols, that is

$$\begin{aligned}\mathcal{A}_\ell \;=\; &\{\text{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z,}\\ &\text{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ø}\}.\end{aligned} \tag{2.1}$$

In a *musical* context − A *sequence* (or *melody* or *melodic line* or *sheet*) is similarly defined as a succession of *notes* from the alphabet. Musical words (*e.g.*, *chords*), sentences and paragraphs can be analogically defined, but will not be necessary in this work. The *note alphabet* $\mathcal{A}_♩$ used this manuscript will be based on the piano, hence including the natural notes and their altered versions, *i.e.*, 12 semi-tones per octave − thus resulting in 88 classes from A_0 to C_8. It is also augmented with a silence token denoted 0.

Interestingly, these two symbolic natures (*e.g.*, text and notes) can be entangled as shown in Figure 2.4. Indeed, this sheet music features a text and singing melody that are expected to "happen" simultaneously in such a way that each part of a word (*i.e.*, a *syllable*) is associated with a note or even hold on several notes in the case of *melisma* − *e.g.*, see "*play?*" on Figure 2.4. It becomes a perspective of research to choose whether to study the text, the notes, or both, which justifies their previous mutual introduction. From a general point of view, both these symbolic information are contained in the sound that a singer would produce. This calls for more expertise on another modality: *orality*.

2.1.3 Voice production and oral modality

Besides symbolic modalities, humans are able to exchange communication messages by generating sounds with their voice and listening to voice sounds from others. The voice is intrinsically speaker-dependent but the production mechanisms of these *oral* messages are shared by all humans. They are therefore briefly detailed in the following.

Acoustic phonetics

The range of ways to articulate sounds when speaking a language is determined by the anatomy of the human body and cannot be expanded upon. This means that all humans have the same potential for making sounds, and sounds from the world's languages rely on similar body configurations. As a result, there exist categories of sounds that can be distinguished based on the way they are produced (*articulation*) or their characteristics (*acoustics*). Each language can use a combination of these sounds to form its own inventory of *phonemes* (Stevens, 2000).

Phonemes are the building blocks of any spoken language as they are defined as the smallest units of sound that can alter the meaning of a word when changed or swapped out, *e.g.*,

$$\begin{array}{lcccc} \text{time} & \mapsto & /\text{t}/ & /\text{a}/ & /\text{i}/ \quad /\text{m}/ \\ \text{dime} & \mapsto & /\text{d}/ & /\text{a}/ & /\text{i}/ \quad /\text{m}/ \end{array}$$

Understanding how phonemes can be *voiced* – and become sounds – requires further information on the vocal apparatus.

Production mechanisms

The *vocal apparatus* is the physical system humans are born with that is capable to turn air flow into sounds. As summarized in Figure 2.5, it can be described by two main elements:

- First, the *voice generator* whose vocal folds, in the case of harmonic sounds, turns air flow from the lungs into periodically spaced pulses, resulting in an excitation with a given tone – or Fundamental Frequency (F0) ;
- Second, the *vocal tract* composed of the tongue, nasal cavity and lips, which is responsible for creating the "color" of the voice – its *timbre*.

Phonemes are classified in two upper categories, namely *vowels* and *consonants*. Consonants (*e.g.*, /t/, /m/) are articulated with complete or partial closure of the vocal tract. Vowels (*e.g.*, /a/, /e/, /i/, /o/, /u/, /y/) are generated without any restriction of the vocal tract, resulting in periodic air pulses leading systematically to harmonic sounds. That is why vowels have a special role in speech and singing analysis: they carry the tone or the melody. In singing, the notes are hold on the vowels of each syllable (Sundberg and Rossing, 1990).

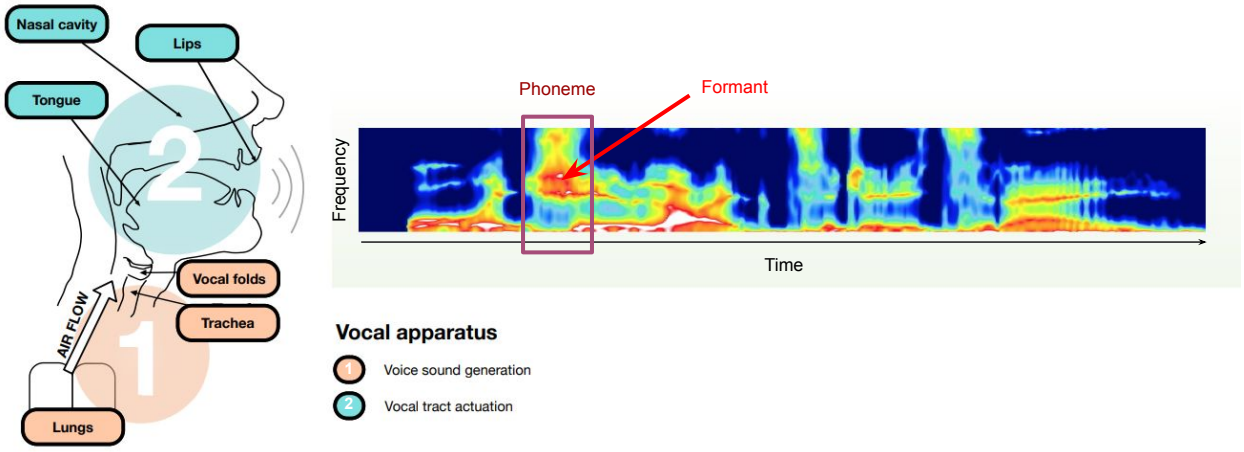


Figure 2.5: Voice production mechanisms. Courtesy of Léane SALAIS (SALAIS et al., 2022).

The form of the vocal tract changes with respect to time as its constituent elements are dynamically in motion. Therefore, in order to capture the time-dependent fluctuations of the phonemes, one must rely on Time-Frequency Representations (TFR) – they will be extensively presented in the next section 2.2.

Formants

The *formants* are the main resonances of the vocal tract. If these resonances are sufficiently dense, formants may be observed on TFR, in regions surrounding the frequency peaks (*partials*) – *i.e.*, maxima of magnitude (energy). Depending on the phonetic context, there can be one or several formants at a given time as depicted in Figure 2.6.

In the case of vowels, two formants are usually detected and analysed. The first formant F_1 is linked to the aperture of the vocal tract while the second formant F_2 is related to the position of the tongue. Singing voice (especially in opera) also has a high third formant F_3 referred to as the *singing formant* and located around 3kHz (SUNDBERG, 1974).

In the case of unvoiced excitation (or unvoiced consonant), only one formant F_1 can be expected as there are no vibrations of the vocal cords generating harmonic structure.

Phoneme identification can be based on formant analysis – *e.g.*, the two first formants F_1 and F_2 have been used to discriminate between vowels by defining the vowel diagram (HELLWAG, 1886). However, formants alone are not sufficient to fully describe voiced phonemes. There are many aspects, characterizing one's expression or interpretation, that can shape the sound of a spoken or sung phoneme. These are the focus of upcoming paragraphs.

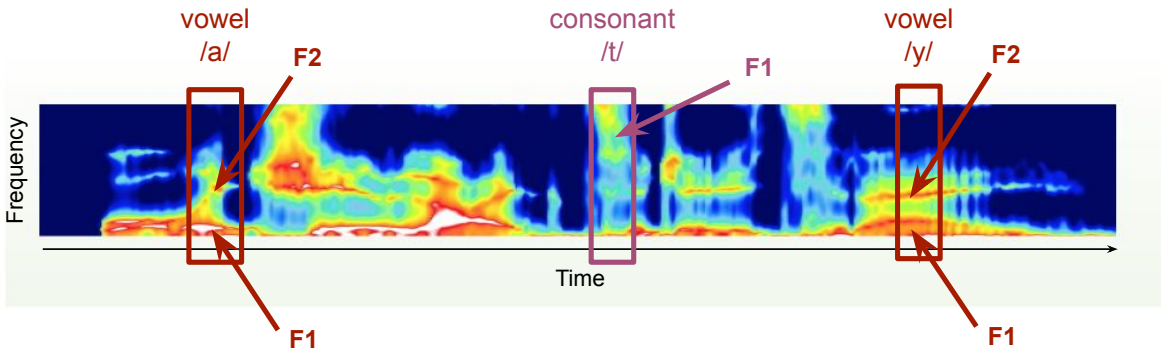


Figure 2.6: Visualization of phonemes and F_1/F_2 formants on Time-Frequency Representation. Courtesy of Léane Salais (Salais et al., 2022).

2.1.4 Expressivity and interpretative style

Humans convey emotions when they talk; singers integrate feelings during a performance. This *expressivity* is precisely what makes a voice sound *natural* and the reason voice goes beyond a simple succession of phonemes and formants. The very simple fact to rise or lower the tone of a voice creates an *intonation* (Collier and Hart, 1975) that shares a state of a mind.

Many studies have been dedicated to emphasize the outstanding ability humans have to communicate social signals when speaking – from *emotions* and their identification (Le Moine and Obin, 2020) to social *attitudes* (Wichmann, 2000)[1] conveyed in speech utterances (Le Moine et al., 2021). Interestingly, analogous phenomena have been highlighted between musicians (Juslin, 2019) for, *e.g.*, humor in music (Rodriguez et al., 2021).

Vocal expressivity is known to induce significant changes in speech characteristics. For instance, the smiling process leads to higher formants F_1 and F_2 during both smiled speech production (Arias et al., 2018b) and perception (Ponsot et al., 2018) to the extent that auditory smiles communicated through formant frequencies can trigger emotional reaction in listeners (Arias et al., 2018a). Facial expression of disgust also modifies the formants (Chong et al., 2018).

In a similar manner, singers carry their very own personal *singing* or *interpretative style* that have driven musicological research in various genres, *e.g.*, French chanson (Chabot-Canet, 2020b; Chabot-Canet et al., 2020) or Rap (Migliore and Obin, 2018). Key aspects, like the *vibrato*, will be mentioned when comparing speech and singing in section 2.2.3.

Like recent collaborative work (Ardaillon et al., 2016), an applicative motivation of this thesis is to uncover some temporal facets of (1) speech production and (2) singing expressivity in a musicological context, and throw them into relief. This will be detailed in section 2.4.

[1] Attitudes are concerned with a speaker's social intention while emotions refer to a speaker's affective state.

Section summary – Human communication

Humans exchange information with one another through *communication messages* that, for a given language, have two main modalities: (1) *symbolic* messages that are dependent on a writing system and an alphabet of symbols $\mathcal{A}$; and (2) *oral* messages that are based on production mechanisms – from the vocal folds generating periodic tones to the vocal tract creating the formants (*timbre* of voiced phonemes). In addition to these physical aspects, voice has a wide range of *expressivity* that comes into play.

2.2 Voice signal processing

The study of a physical phenomenon can be described as one or several quantities, *i.e.*, physical variable(s) depending on *time* and/or *space*. Extracting information from the observations at our disposal often requires to look closely at these temporal or spatial aspects. It is common to refer to a temporal evolution as *signal* and to a spatial evolution as *image*.

In this context, signal processing is a wide scientific field (Rabiner and Gold, 1975) dedicated to extract, manipulate and transform information contained in signals. While there exist many natures of signals (*e.g.*, electric measurements, temperature values, weather maps, etc.), the research in this thesis is exclusively focused on sounds. Indeed, voice, audio and music are sound signals and, as such, can be manipulated with signal processing tools (Müller, 2015).

This section naturally exposes the main notions necessary to study voice and audio from a signal processing perspective. Once the basics from digital audio are presented, the main characteristics – or *features* – one can extract from voice signals are defined. Since voice can result in speech or singing signals that share similarities but also differences, a comparison between speaking and singing voices is drawn. Finally, two major signal-oriented research fields highly tied to each other are presented: analysis and synthesis of voice signals.

2.2.1 Digital audio basics

The starting point of (sound) *signal* processing is the *signal* itself. When a sound is produced in a fluid, the surrounding molecules are set in motion by the sound waves, creating pressure variations step by step in both space and time.

A sound signal $\int$ is therefore a function varying *continuously* in *time* that maps an instant $t_c \in \mathbb{R}$ to a vibration amplitude $\int(t_c)$, that is

$$\begin{array}{rccc} \int & : & \mathbb{R} & \rightarrow & \mathbb{R} \\ & & t_c & \mapsto & \int(t_c). \end{array} \tag{2.2}$$

From continuous-time to discrete-time signals

In practice, sound signals are captured and often emitted by systems such as microphones and loudspeakers that cannot deal with continuous values for time and amplitudes as it would require an infinite storage capacity. Therefore, in the context of *digital* signal processing, audio signals are *discretized* by means of *quantization* and *sampling*.

Quantization consists in limiting the amplitude range of $\int$ to a finite number of allowed values $\mathcal{Q}_b$, depending on the number of digital bits N_b implied in the process, such that each amplitude $\int(t_c)$ becomes the closest permitted value $\int_q(t_c) \in \mathcal{Q}_b$, *i.e.*,

$$\int_q(t_c) = \min_{q \in \mathcal{Q}_b} |\int(t_c) - q|. \tag{2.3}$$

Sampling consists in measuring the signal amplitude at only discrete instants t_n with $n \in \mathbb{N}$. In the most classical sampling setup for audio, known as equidistant sampling, two successive discrete instants are separated by a constant, fixed *sampling period* T_s so that $t_n = nT_s$.

This leads to define the *discrete*-time signal x from the continuous-time signal $\int$ as follows:

$$\begin{array}{rcccl} x & : & \mathbb{N} & \to & \mathcal{Q}_b \\ & & n & \mapsto & x[n] \quad = \int_q(t_n) \quad = \min\limits_{q \in \mathcal{Q}_b} |\int(nT_s) - q|. \end{array} \tag{2.4}$$

Frequency and harmonic analysis

A signal is referred to as *periodic* if there exists a minimal quantity called *period* $T_0 \in \mathbb{R}_+^*$ such that $\forall t \in \mathbb{R}, \int(t) = \int(t + T_0)$. The *(fundamental) frequency* F_0 is the reciprocal of the period:

$$F_0 = 1/T_0. \tag{2.5}$$

If the period (in sec.) defines the duration before the signal repeats itself, the frequency (in HERTZ, Hz) measures the number of repetitions in an unit time interval. Integer multiples of the fundamental frequency are known as *harmonics*.

The *harmonic decomposition* of signals, which is the basis of all FOURIER's analysis theory (HIGGINS et al., 1996), has revealed that any signal (even non-periodic ones) can actually be represented by a summation/integral of periodic signals. Hence, a signal contains information at many frequencies − from zero to an upper limit $F_{\max}$ above which no information is left.

It was shown that discretizing a signal involves a sampling period T_s, one can thus define similarly the *sampling frequency*:

$$F_s = 1/T_s. \tag{2.6}$$

One major result of digital signal processing is the NYQUIST-SHANNON-KOTELNIKOV theorem stipulating that a signal whose maximum frequency is $F_{\max}$ can be sampled without loss of information if and only if the sampling frequency verifies the SHANNON's criterion:

$$F_s > \frac{F_{\max}}{2}. \tag{2.7}$$

As far as humans are concerned, their hearing system has a frequency upper bound around 20kHz and voice components, whether voiced or unvoiced, are rather small for frequencies above 8kHz. As a result, given Eq. (2.7) usual values for the sampling frequency F_s are $44,1$Hz or 48kHz for the recording industry and 16kHz for voice processing technologies. One exception to mention is the reduced bandwidth in telephony (less than 4kHz), for which a sampling frequency of 8kHz is enough.

Due to the mentioned harmonic decomposition, it is extremely common to work in the frequency domain rather than the time domain as many operations, such as *filtering* (*i.e.*, alteration of the content of a signal), are easier to perform in the former one. Therefore, the most relevant frequency representations of audio signals are further introduced.

Discrete Fourier Transform (DFT)

The Discrete Fourier Transform (DFT) is a mathematical operator allowing to decompose a discrete-time signal into its constituent frequencies, *i.e.*, its *spectrum* or *spectral representation*. The DFT of the discrete-time signal x of length N – denoted $\text{DFT}[x]$ – is defined as follows:

$$\begin{array}{rcccl} \text{DFT}[x] & : & \mathbb{N} & \to & \mathbb{C} \\ & & f & \mapsto & \text{DFT}[x](f) = \displaystyle\sum_{n=0}^{N-1} x[n] \exp\left(-\frac{2j\pi fn}{N}\right) \end{array} \tag{2.8}$$

where $j = \sqrt{-1}$ is the imaginary unit and $f \in \{0, \dots, F-1\}$ is the f^{th} frequency. The Fourier coefficient $\text{DFT}[x](f) \in \mathbb{C}$ contains the magnitude and the phase of the sinusoidal component of the signal with frequency fF_s/N.

Note that a common variant of the Discrete Fourier Transform (DFT) is the Discrete Cosine Transform (DCT) that makes use of a cosine function instead of a complex exponential in Eq. (2.8). In the case of DCT, Fourier coefficients are real numbers.

Short-Time Fourier Transform (STFT)

The spectral content of voice or audio recordings is evolving with respect to time, *e.g.*, people do not talk permanently with the exact same "tone" and chords/notes usually vary in songs. Thus, sound signals are intrinsically *non-stationary* as their statistical properties are time-dependent.

A single Discrete Fourier Transform (DFT) is not sufficient to study such signals as it would result in a representation that would not capture the temporal evolution of the spectral properties. Time-Frequency Representations (TFR) are much more relevant features to manipulate in this regard.

The Short-Time Fourier Transform (STFT) is specifically designed for the analysis of non-stationary signals as it considers them as successive short signals on which local DFTs are computed and concatenated.

This is done by sliding a temporal *window* w of fixed sample size W, step by step with overlapping. Each resulting sub-signal is referred to as a *frame* and indexed through time by $t \in \mathbb{N}$. The sliding and overlapping parameter is known as *hop size* H and typically defines the *temporal precision* $\delta t = H/F_s$. All in all, the STFT $\mathbf{x}$ of the signal x is a function of both time frame t and frequency f:

$$\begin{array}{rcccl} \mathbf{x} & : & \mathbb{N} \times \mathbb{N} & \to & \mathbb{C} \\ & & [t, f] & \mapsto & \mathbf{x}[t, f] = \sum_{n=0}^{N-1} w[n] x[n + tH] \exp\left(-\frac{2j\pi f n}{N}\right). \end{array} \tag{2.9}$$

Windowing the signal might prevent edge effects like the Gibbs phenomenon (Gottlieb and Shu, 1997) from appearing but distort the samples' weight, which is compensated by overlapping. STFT computations classically rely on the Hann (or Hanning) window:

$$w[n] = \begin{cases} \frac{1}{2}\left(1 - \cos\left(2\pi \frac{n}{W}\right)\right) & \text{if } n \in \{0, \dots, W-1\} \\ 0 & \text{elsewhere.} \end{cases} \tag{2.10}$$

Spectrogram

The *spectrogram* $|\mathbf{x}|$ is the magnitude of the STFT, that is

$$\begin{array}{rcccl} |\mathbf{x}| & : & \mathbb{N} \times \mathbb{N} & \to & \mathbb{R} \\ & & [t, f] & \mapsto & |\mathbf{x}|[t, f] = \left| \sum_{n=0}^{N-1} w[n] x[n + tH] \exp\left(-\frac{2j\pi f n}{N}\right) \right|. \end{array} \tag{2.11}$$

It is one of the most used representations for audio and music processing.

Convolution & signal filtering

Convolution, denoted $*$, is an extremely popular operation as it mathematically embodies linear *filtering*, *i.e.*, the modification of some inputs by application of a *kernel* $\mathcal{K}$. It concretely performs a smooth weighted average between the kernel and the inputs at each point.

In the case of a 1-D convolution − inputs are typically time series like a raw signal x. The kernel $\mathcal{K}$ is also one-dimensional and the resulting filtering is expressed as:

$$(x * \mathcal{K})[n'] = \sum_n x[n] \mathcal{K}[n' - n]. \tag{2.12}$$

In the case of a 2-D convolution − inputs are typically images or, for audio, any TFR like the spectrogram $|\mathbf{x}|$. The kernel $\mathcal{K}$ is therefore two-dimensional and the resulting filtering is expressed as:

$$(|\mathbf{x}| * \mathcal{K})[t', f'] = \sum_t \sum_f |\mathbf{x}|[t, f] \mathcal{K}[t' - t, f' - f]. \tag{2.13}$$

2.2.2 Voice features

The general notions and frequency representations of sound signals have been introduced, allowing one to specifically define the main characteristics – or *features* – related to voice signals. In order to clearly emphasize the relationships among these features, their Time-Frequency Representations (TFR) are summarized in Figure 2.7.

Pitch – Fundamental Frequency (F0)

The *pitch* of a voice, similar to the pitch of a note, is related to its perceived highness or lowness. The common auditory sensation is that the greater the pitch, the higher the voice. It allows to naturally distinct voice tones from one another and order them in a gradually increasing scale from low to high.

In practice, when a given pitch is perceived, it can be mapped to the frequency of a pure sound (sine wave). Therefore, throughout this manuscript, the Fundamental Frequency (F0) will be the main feature to characterize a pitch and both terms may be used interchangeably, although there is more to pitch than a single F0 value, *e.g.*, (Révész, 1954; Shepard, 1982).

Log-Mel-spectrogram

Psychoacoustic studies dedicated to human auditory system revealed that human hearing perception scales up logarithmically in frequency and intensity (Fastl and Zwicker, 2006). The ear, indeed, is able to distinguish (1) very calm from very loud sounds; and (2) lower from higher pitches with a better discriminative pitch perception in low frequencies. These considerations are not taken into account in a simple spectrogram as defined in Eq. (2.11).

Regarding the ear's loudness perception, it is more relevant to compute log-scaled magnitudes than linear ones. By means of the operation $|\mathbf{x}| \mapsto 20\log_{10}(\epsilon + |\mathbf{x}|)$, with ϵ a small value, the spectrogram is turned into a *log-spectrogram*.

Regarding the ear's frequency perception, Stevens et al. (1937) introduced a perceptive "mel"ody scale – the *Mel scale* – specifically designed to represent frequencies in line with human audition:

$$M(f) \quad = \quad 2595 \times \log_{10}\left(1 + \frac{f}{700}\right). \tag{2.14}$$

Mapping the spectrogram linear frequencies to the Mel scale is usally done with F overlapping triangular filters referred to as *Mel filterbank*. In the end, the *mel-spectrogram* is obtained.

Coupling both loundness and frequency aspects, one can compute the *log-mel-spectrogram*, which contains all perceptual features, making it a robust default representation of voice and audio signals.

Spectral envelope

A classical audio production model is the source-filter model, which assumes that a signal $\int$ results from an original excitation (the *source*) e modified by a resonator (the *filter*) r. Remembering section 2.1.3, the voice correlates are the vocal cords for the source and the vocal tract for the filter. For each time $t_c \in \mathbb{R}$, the signal is therefore expressed as a temporal convolution:

$$\int(t_c) \quad = \quad (e * r)(t_c). \tag{2.15}$$

In the frequency domain, denoting $\mathcal{F}$ a general FOURIER transform, the spectrum are multiplied,

$$\mathcal{F}[\int](f) \quad = \quad \mathcal{F}[e](f) \quad \times \quad \mathcal{F}[r](f). \tag{2.16}$$

Taking the logarithm of Eq. (2.16), one can isolate spectral components, respectively the pitch (from the *source*) and the harmonic content or *timbre* (resonances/attenuations from the *filter*). Summing up, the *cepstrum* is defined as:

$$\log\left(\mathcal{F}[\int](f)\right) \quad = \quad \underbrace{\log\left(\mathcal{F}[e](f)\right)}_{\text{fast variations in } f} \quad + \quad \underbrace{\log\left(\mathcal{F}[r](f)\right)}_{\text{slow variations in } f}. \tag{2.17}$$

The (log-)*spectral envelope* precisely corresponds to the resonator term $\log\left(\mathcal{F}[r]\right)$. It is supposed to be a smooth and F0-free representation carrying only information from the vocal tracts (voice formants and timbre). The problem is r is concretely unknown. In practice, a Discrete Cosine Transform (DCT) of the whole cepstrum is performed to separate high and low *quefrencies* and isolate the specral envelope from the excitation structure. See (RÖBEL and RODET, 2005) for another method for spectral envelope estimation.

Mel-Frequency Cepstral Coefficients (MFCCs)

The Mel-Frequency Cepstral Coefficients (MFCCs) have been one of the widespread features in the audio and voice literature, allowing a compact representation of above-mentioned spectral envelope (DAVIS and MERMELSTEIN, 1980).

The MFCCs are the low order coefficients (*i.e.*, , usually up to 13) from the Discrete Cosine Transform (DCT) calculated on the log-mel-spectrogram. According to previous source-filter modeling, MFCCs are expected to globally follow the spectral envelope and be pitch-independent to a certain extent.

Phonetic characterization

As discussed in section 2.1.3, formants are induced by the vocal tract. Therefore, as previously exposed, spectral envelope and/or MFCCs features – given their strong links – seem particularly well-suited for representing formants and phonetic information. Concretely, a given spectral envelope should be enough to recognize a precise phoneme and reciprocally.

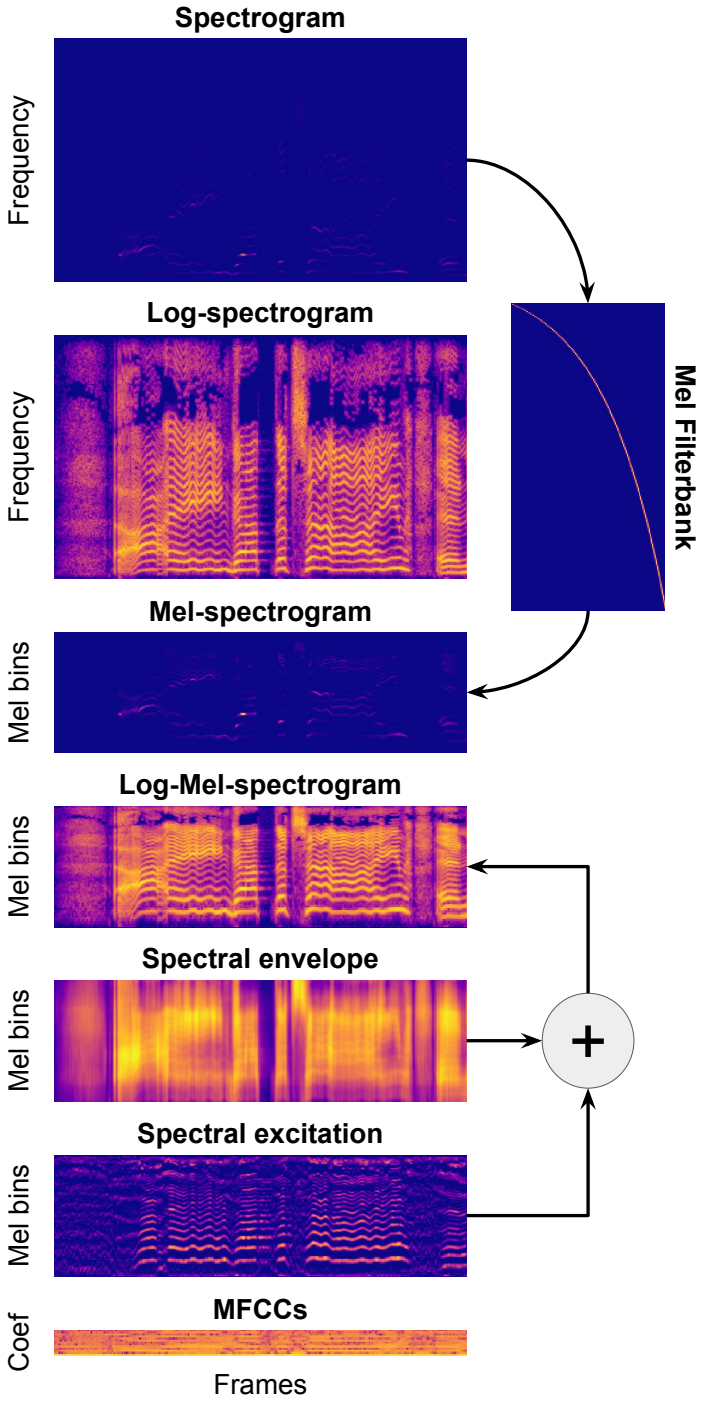


Figure 2.7: Time-Frequency Representations (TFR) in voice signal processing. Example is sung vocals "I ain't got the ***time***" from Amy WHINEHOUSE's Rehab, HANSEN (2012)'s dataset.

2.2.3 Speech *vs* singing comparison

The main representations used to study and process voice signals have been presented. While these features can be exploited for both speech and singing, it is undeniable that – despite common characteristics – speaking and singing voices cannot be classified as equals. This section provides insights on these differences that are worth mentioning. A visual example highlighting these variations can be found in Figure 2.8.

Either specialized on singing (Sundberg and Rossing, 1990) or speech (Kent and Read, 2002), or clearly addressing their comparison (Livingstone et al., 2013; Ohishi et al., 2005), even up to the recording, release and analysis of a specific dataset (Duan et al., 2013), there is a long research history dedicated to such questions. Based on these references, five main differences between speaking and singing voices are listed. They are either related to characteristics of the speaker's vocal apparatus (see section 2.1.3), or prosodic gestures and expressivity (see section 2.1.4). A summary of these speech/singing discrepancies is proposed:

- *Vocal tract shape* – During the generation of voice signals, the shape of the vocal tract (*i.e.*, mouth, tongue, and throat) varies to produce the several phonemes and sounds. A first difference between speech and singing voice is that the vocal tract changes much faster in speech than in singing.

- *Intonation* – Then, *intonation*, is an important way to convey meaning and emotion through voice by, *e.g.*, slightly adapting F0 curves. There are generally faster variations in speech but they may purposely be exaggerated in a singing context, while rather subtle in spoken utterances.

- *Vibrato* – The *vibrato* may be the prime example of speech-singing differences as this effect is not observed for spoken voice. Created by a fine control of the muscles of the vocal apparatus, the vibrato manifests through local and rapid fluctuations in the pitch of a note. It has been highly analyzed by musicologists to understand a singer's style (Chabot-Canet et al., 2020) as it is a major piece of expressivity and is ubiquitous in, *e.g.*, French chanson or popular music.

- *Fundamental Frequency (F0)* – The fundamental frequency of spoken voice typically ranges between 80Hz and 200Hz. The singing voice has a much wider tessitura as sung pitches can reach high frequencies, up to 1kHz and more.

- *Phoneme duration* – In speech, phoneme duration is around 20-200ms (as measured on (Zue et al., 1990)). In singing, for which a pitched note is associated with a vowel, phonemes can be hold much longer as singers tend to sustain vowels for longer periods. In doing so, they can add expressivity to their performances, duration being in itself an expressive means, via intensity variations over a syllable or vocal techniques mastering.

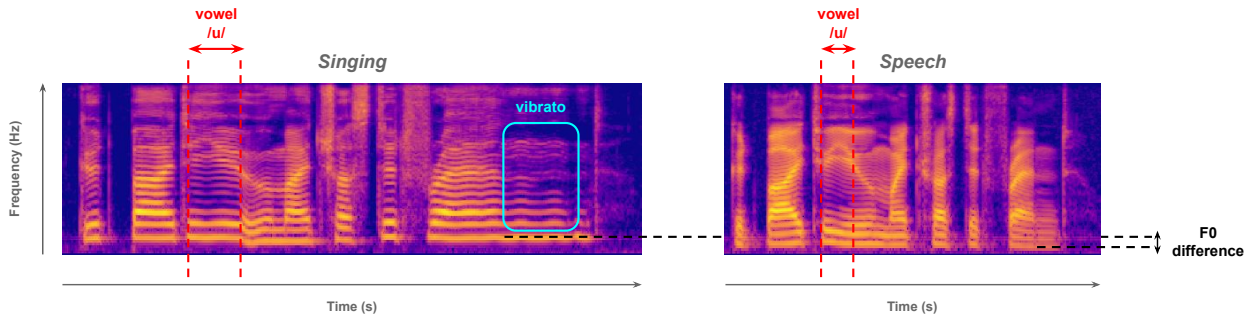


Figure 2.8: Illustration of differences between speech and singing: vowel duration, F0, vibrato. Example is the same sentence read and sung by KENN_17 from Duan et al. (2013)'s dataset.

Moving forward, having covered the diversity encountered in voice signals – both in terms of their features and their quality (speech *vs* singing) – the core tasks and challenges involved in current voice research can be exposed: analysis and synthesis.

2.2.4 Voice analysis and synthesis

Scientists in voice-related research have extensively worked on three main fields – voice *analysis*, *synthesis* and *transformation*. The later field, voice transformation, consists in modifying one or some feature(s) of a given voice, *e.g.*, changing the attitude in a recording (Tao et al., 2006) or convert several vocal attributes (Sisman et al., 2020), but it is out of the scope of this thesis. On the contrary, the two former, synthesis and analysis, have an important role in this work.

Synthesis of voice signals

Voice synthesis is the field aiming at producing artificial, yet plausible, voices. This includes synthesis of both speech and singing signals that should be perceived as natural as possible. As in many other voice-oriented tasks, early research was focused on speech and the same procedures were later extended and adapted to singing.

The first attempt to synthesize speech dates back to 1791 with the "speaking machine" (von Kempelen, 1791). This mechanical system, and others that followed (Dudley et al., 1939), was designed to mimic the physical mechanisms of speech production that were introduced in section 2.1.3. Far from this original system that required manual intervention, speech synthesis has benefited from the progress in computer sciences towards the rise of automatic methods.

The modern era of speech synthesizer has been marked with the development of Text-to-Speech (TTS) systems, that are capable to directly synthesize acoustic signals with relevant speech parameters for a given text to be voiced. Most current systems rely on approaches based on artificial intelligence for both speech (Ning et al., 2019) and singing (Cho et al., 2021) synthesis. Related deep learning notions for voice will be covered in section 3.1.

Today, voice synthesis is ubiquitous in daily life technologies and has found relevance in telecommunications (*e.g.*, vocal assistants (Pal et al., 2019)), medicine (Brunow and Cullen, 2021; Řepová et al., 2021) or artistic creations (*e.g.*, video games (Farner et al., 2008)), etc.

While this thesis is not dedicated to voice synthesis, there exists a synthesis method that resonates with this work. It is *concatenative synthesis* that achieves TTS by creating smooth transitions between phonemes – or groups of successive phonemes (*biphones*, *triphones*) – from pre-existing voice recordings (Ardaillon, 2017). This approach intrinsically requires a strong correspondence between audio and text, which is in line with this research. Thus, voice synthesis will remain an interesting application that will be discussed in section 2.4.

Analysis of voice signals

Voice analysis is the field aiming at extracting all kind of information from voice signals, which covers a wide scope of possible objectives. Among numerous examples, typical voice analysis tasks are concerned with pitch estimation (Ardaillon and Roebel, 2019), language and speaker identification (Tirumala et al., 2017), keyword spotting (López-Espejo et al., 2021), query by singing and humming (Liang et al., 2021a), or vocals extraction (Cohen-Hadria et al., 2019). There is also an active research field – Music Information Retrieval (MIR) – specialized on similar considerations for music (Müller, 2007b; Schedl et al., 2014).

Early research was based on *hand-crafted* methods, *i.e.*, deriving systematic rules upon computation of audio descriptors (Peeters et al., 2011). With the emergence of large collections of data and higher computational power over the years, recent research rather relies on *data-driven* methods (*e.g.*, machine learning and deep learning) that are able to construct relevant prediction patterns from data themselves (Papakostas et al., 2017), or a coupling of both. The required background to manipulate data-driven systems is introduced in the next chapter.

This thesis inscribes itself in voice analysis as it aims to understand and uncover the underlying relationships between voice and symbolic representations. Hence, the upcoming section is dedicated to the tasks intimately related to this research: transcription and alignment.

Section summary – Voice signal processing

Voice signal processing a wide scientific field allowing to define, represent and transform information in *digital voice signals*. These signals result from a temporal sampling with a usual *sampling frequency* of F_s =16kHz. Among the numerous existing *voice features*, the *log-mel-spectrogram* and *spectral envelope* are two Time-Frequency Representations (TFR) of highest interest. The later captures the formants and phonemes while the former integrates all human perception aspects. Based on TFR, active research on voice *analysis* and *synthesis* has been conducted. In practice, despite the known speech-singing differences (*e.g.*, vowels duration, etc.), speech approaches tend to be adapted to singing.

2.3 Transcription and alignment tasks

Transcription and alignment are two prominent analysis tasks in many fields. They have gained tremendous interest and found major and diverse implications in the audio domain, and especially in the voice community, and beyond. The research presented in this manuscript is dedicated to voice-to-symbols synchronization, yet, given the high proximity between alignment and transcription in terms of concepts, formalization, modeling approaches as well as general multi-modal problems, both of them will be studied and referred to throughout the entire document.

2.3.1 Definitions

The common definitions and notations that will used for both tasks are introduced first. These concepts are, more generally, in line with all literature processing data *sequences* − a *sequence* being a finite succession of elements, with a notion of order through an indexing via indices.

Sequence modeling & decoding

Let $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{H}$ denote three feature spaces. Let $\mathscr{X}^*$, $\mathscr{Y}^*$ and $\mathscr{H}^*$ be the sets of all sequences over $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{H}$, respectively.

Let m denote a message emitted in two different forms with, *e.g.*, different modalities and/or expressivity. Let $\mathbf{x} \in \mathscr{X}^*$ denote the first emission and $\mathbf{y} \in \mathscr{Y}^*$ denote the second emission. Let T be the length of $\mathbf{x}$ such that $\mathbf{x} = \{\mathbf{x}_0, \dots, \mathbf{x}_{T-1}\}$ with $\forall t \in \{0, \dots, T-1\}, \mathbf{x}_t \in \mathscr{X}$. Let M be the length of $\mathbf{y}$ such that $\mathbf{y} = \{\mathbf{y}_0, \dots, \mathbf{y}_{M-1}\}$ with $\forall m \in \{0, \dots, M-1\}, \mathbf{y}_m \in \mathscr{Y}$.

Transcription and alignment tasks can be seen as twofold procedures composed of

- A similar modeling function $\mathcal{M}$, such that

$$\begin{array}{rccc} \mathcal{M} & : & \mathscr{X}^* & \rightarrow & \mathscr{H}^* \\ & & \mathbf{x} & \mapsto & \mathcal{M}(\mathbf{x}). \end{array} \tag{2.18}$$

  Applied on the first emission $\mathbf{x}$, this function returns the quantity $\mathcal{M}(\mathbf{x})$, which has a different designation (*e.g.*, saliency map, posteriorgram) according to the task objective. With no loss of generality, it can be called *latent code*, *hidden code* or *hidden representation*.

- A specialized decoding module $\mathcal{D}$, which differs between transcription and alignment tasks, but somehow connects the two emissions $\mathbf{x}$ and $\mathbf{y}$.

The existing and chosen approaches for designing the modeling and decoding functions are at the core of this research and will be mathematically exposed in the following chapters.

Sequence similarity & alignment

The notion of *alignment* comes from the long research history dedicated to the relationships and similarities between sequences, which manipulates specific tools that are introduced hereinafter.

To define a notion of correspondence between two sequences $\mathbf{x} \in \mathscr{X}^*$ and $\mathbf{y} \in \mathscr{Y}^*$, a sequence $\boldsymbol{\pi}$ of length $K = \max(M, T)$ of ordered tuples over the indices of $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$\boldsymbol{\pi} = \{\boldsymbol{\pi}_k\} \quad \text{with} \quad \forall k \in \{0, \dots, K-1\}, \quad \boldsymbol{\pi}_k = (t_k, m_k) \in \{0, \dots, T-1\} \times \{0, \dots, M-1\} \tag{2.19}$$

This sequence goes by many names including a *path* or *pathway* (Needleman and Wunsch, 1970), *warping* or *warping function* (Itakura, 1975), or *alignment path*. The later option will be chosen in this work. The ensemble of all paths between elements of $\mathscr{X}^*$ and $\mathscr{Y}^*$ is denoted $\Pi(\mathscr{X}^*, \mathscr{Y}^*)$. There exist several properties – or alignment constraints – that a given path $\boldsymbol{\pi} \in \Pi(\mathscr{X}^*, \mathscr{Y}^*)$ is expected to satisfy. Namely,

- *Boundary conditions* such that the alignment starts (resp. ends) at the first (resp. last) indices of the two sequences:

$$\boldsymbol{\pi}_0 = (0, 0) \quad \text{and} \quad \boldsymbol{\pi}_{K-1} = (T-1, M-1). \tag{2.20}$$

- *Monotony*, which stipulates that the path can only move forward, such that successive indices verify:

$$\forall k \in \{0, \dots, K-2\}, \quad t_k \leq t_{k+1} \quad \text{and} \quad m_k \leq m_{k+1}. \tag{2.21}$$

- *Limited progression* of the path, such that only some transitions are permitted between successive states $\boldsymbol{\pi}_k$ and $\boldsymbol{\pi}_{k+1}$ according to a set of allowed gaps $\mathcal{G}$, *i.e.*,

$$\forall k \in \{0, \dots, K-2\}, \boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k = (t_{k+1} - t_k, m_{k+1} - m_k) \in \mathcal{G}. \tag{2.22}$$

The Figure 2.9 illustrates an alignment verifying these constraints between audio and text (*e.g.*, $\mathbf{x}$ represents T spectral frames and $\mathbf{y}$ represents M characters) when $\mathcal{G} = \{(1, 0), (1, 1)\}$.

Finally, let $\mathscr{S} : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ be a similarity measure between the elements of $\mathscr{X}$ and $\mathscr{Y}$. The *alignment* between the sequences $\mathbf{x} \in \mathscr{X}^*$ and $\mathbf{y} \in \mathscr{Y}^*$ corresponds to the *optimal path* $\boldsymbol{\pi}^*$, that is the path maximizing the accumulative similarity measures (*i.e.*, for all tuples of indices). It reads, with $\mathbf{x}_t \in \mathscr{X}$ and $\mathbf{y}_m \in \mathscr{Y}$ the tth and mth elements of $\mathbf{x}$ and $\mathbf{y}$, respectively,

$$\boldsymbol{\pi}^* \quad = \quad \underset{\boldsymbol{\pi}}{\operatorname{argmax}} \sum_{(t,m) \in \boldsymbol{\pi}} \mathscr{S}(\mathbf{x}_t, \mathbf{y}_m). \tag{2.23}$$

The resulting *alignment score* $\mathbf{A}^* \in \mathbb{R}$ between the sequences is the full similarity measure of this optimal path:

$$\mathbf{A}^* \quad = \quad \sum_{(t,m) \in \boldsymbol{\pi}^*} \mathscr{S}(\mathbf{x}_t, \mathbf{y}_m). \tag{2.24}$$

Solving Eq. (2.23) is a typical *optimization* problem. Techniques for efficiently computing this path via Dynamic Programming (DP) will be seen in the next chapter – in section 3.3.1.

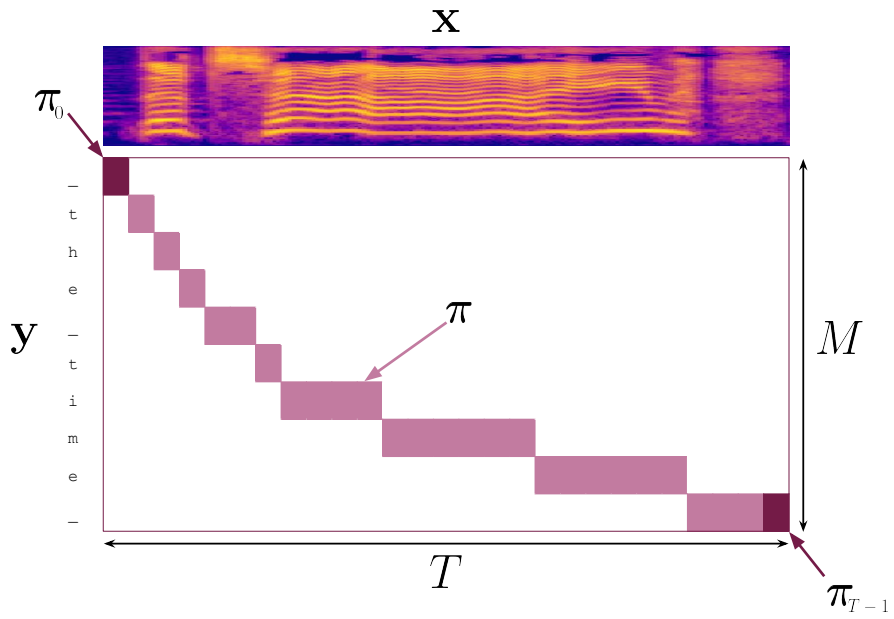


Figure 2.9: An alignment path $\boldsymbol{\pi}$ between sequences $\mathbf{x}$ (voice log-mel-spectrogram) and $\mathbf{y}$ (text).

Transcription/Recognition

Transcription – or *recognition* – is the task that aims at predicting a symbolic sequence of a message from some input data representing its emission. Concretely, from the message emission in a given modality $\mathbf{x} \in \mathscr{X}^*$, it is expected to retrieve $\mathbf{y} \in \mathscr{Y}^*$ – where in practice, $\mathscr{Y}$ is a finite alphabet of symbols $\mathcal{A}$. The Figure 2.10 illustrates the philosophy of this task.

Although there might be a little distinction between transcription and recognition terms[2], they will be both used interchangeably from now on, very much as done in the literature.

In a transcription context associated with a modeling $\mathcal{M}$, the decoding module $\mathcal{D} \equiv \mathcal{D}_{\mathrm{T}}$ becomes

$$\begin{array}{rccccc} \mathcal{D}_{\mathrm{T}} & : & \mathscr{H}^* & \to & \mathscr{Y}^* & \\ & & \mathcal{M}(\mathbf{x}) & \mapsto & (\mathcal{D}_{\mathrm{T}} \circ \mathcal{M})(\mathbf{x}) & = \hat{\mathbf{y}} \end{array} \tag{2.25}$$

and therefore estimates a sequence $\hat{\mathbf{y}} \in \mathscr{Y}^*$ as close as possible to the real sequence $\mathbf{y}$.

Classical options for the decoding module $\mathcal{D}_{\mathrm{T}}$ are greedy search (Germann, 2003) or beam search decoding (Freitag and Al-Onaizan, 2017) that are designed to estimate the most probable sequence given an hidden representation like $\mathcal{M}(\mathbf{x})$.

[2]Actually, recognition is a specific and constrained case of transcription that aims to predict one-to-one correct symbol while transcription, more generally, aims to reflect a correct content (as a semantic concept), which is more flexible than per-symbol prediction.

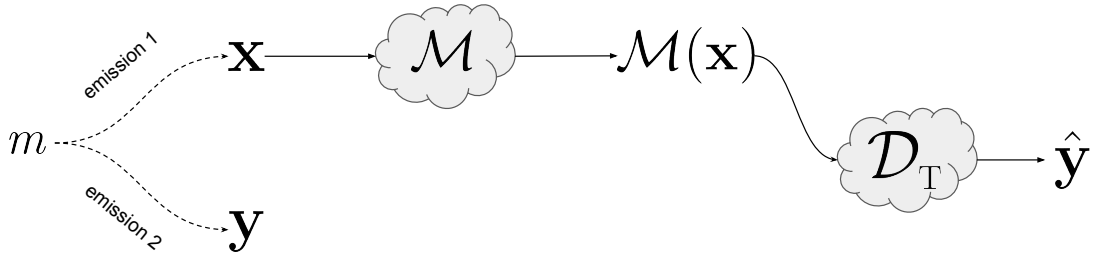


Figure 2.10: Transcription task overview: an emission $\mathbf{x}$ of a message m is processed by a model $\mathcal{M}$ prior to the estimation $\hat{\mathbf{y}}$ of a second and symbolic emission $\mathbf{y}$ thanks to a decoder $\mathcal{D}_{\mathrm{T}}$.

Alignment/Synchronization

Alignment – or *synchronization* – is the task that consists in finding the precise correspondence between several emissions of a message. The events to tie, hence, share a common underlying information but differ in their respective emissions due to different modalities and/or expressivity. Concretely, from two emissions $\mathbf{x} \in \mathscr{X}^*$and $\mathbf{y} \in \mathscr{Y}^*$ of the same message m, it is expected to retrieve the optimal path $\boldsymbol{\pi}^* \in \Pi(\mathscr{X}^*, \mathscr{Y}^*)$ (and its similarity score $\mathbf{A}^* \in \mathbb{R}$) between them. The Figure 2.11 illustrates the philosophy of this task.

According to its Latin etymology *linea* – meaning *line* or *string* – the word "alignment" intrinsically carries the idea to draw a line between objects or concepts and, by doing so, mapping them out. *Synchronization* is the specific case of alignment for which *time* is involved in the process, *i.e.*, when linking the different events requires and is done with a temporal reference – a *timeline*. For such cases, one can refer synonymously to both *alignment* and *synchronization* terms, as it will be done in the rest of the writing.

In this thesis, more specifically, *forced alignment* problems are tackled, as finding the best path ensures that none of the events to align is missed. Indeed, the above-mentioned constraints, *i.e.*, Eq. (2.20), Eq. (2.21) and Eq. (2.22), make sure that each event in the sequence $\mathbf{y}$ has at least one corresponding event in the other sequence $\mathbf{x}$ (admitting that $M \leq T$ otherwise some symbol from $\mathbf{y}$ would be left out). Many alignment issues actually deal with forced alignment.

In such a context associated with a modeling $\mathcal{M}$, the decoding module $\mathcal{D} \equiv \mathcal{D}_\pi$ becomes

$$\begin{array}{rcccccc} \mathcal{D}_\pi & : & \mathscr{H}^* \times \mathscr{Y}^* & \rightarrow & \Pi(\mathscr{X}^*, \mathscr{Y}^*) & & \\ & & (\mathcal{M}(\mathbf{x}), \mathbf{y}) & \mapsto & \mathcal{D}_\pi\left(\mathcal{M}(\mathbf{x}), \mathbf{y}\right) & = & \hat{\boldsymbol{\pi}}^* \end{array} \tag{2.26}$$

and therefore estimates an alignment path $\hat{\boldsymbol{\pi}}^*$ as close as possible to the true optimal path $\boldsymbol{\pi}^*$.

As a reminder, the modeling function $\mathcal{M}$ shares a common role between transcription and alignment: to encode the input data $\mathbf{x}$ into a more relevant and exploitable hidden representation before decoding. This function has been implemented with multiple approaches over the years that will be exposed in the upcoming chapter covering mathematical essentials.

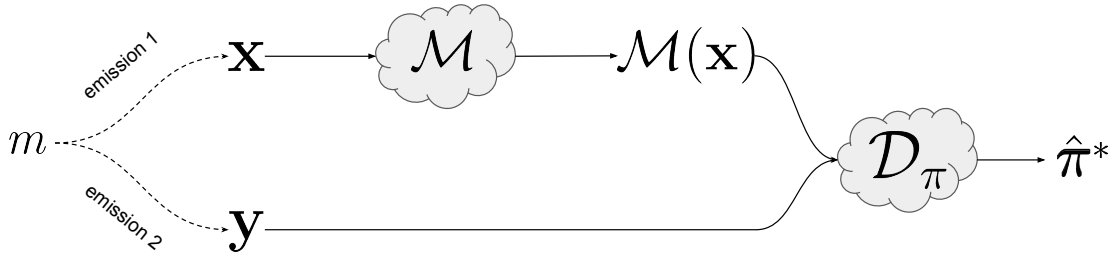


Figure 2.11: Alignment task overview: an alignment path $\hat{\boldsymbol{\pi}}^*$ between two emissions $\mathbf{x}$ and $\mathbf{y}$ of a message m is estimated through a modeling function $\mathcal{M}$ and a mixed decoder $\mathcal{D}_\pi$.

2.3.2 Transcription algorithms

In this section, the wide variety of transcription algorithms are presented starting from general then audio-related transcriptions and with a final emphasis on the voice transcription tasks. Some of the mentioned examples are shown in Figure 2.12.

General transcription

There are many applications of transcription algorithms in various fields such as Computer Vision (CV) or Natural Language Processing (NLP) .

Concrete examples are the optical recognition of characters (Srivastava et al., 2022) or handwriting digits (LeCun et al., 1989) in images, or address numbers from street views (Goodfellow et al., 2014).

From videos, one can attempt to transcribe sign language (Bantupalli and Xie, 2018) or identify movements in sports (Cust et al., 2019), which are special cases of automatic gesture recognition (Wu and Huang, 1999).

Through their application in studies of texts and languages, transcription algorithms also have their impacts in NLP (Dister et al., 2009). One core domain is Machine Translation (MT) which proceeds to text-to-text predictions by automatically translating content from a source language to a target language (Stahlberg, 2020).

Audio transcription

The audio domain, closer to this work, also requires transcription algorithms for diverse tasks. Besides speech/singing recognition, which are discussed below, a main research objective is to recover symbolic music (*e.g.*, music sheet) from audio recordings.

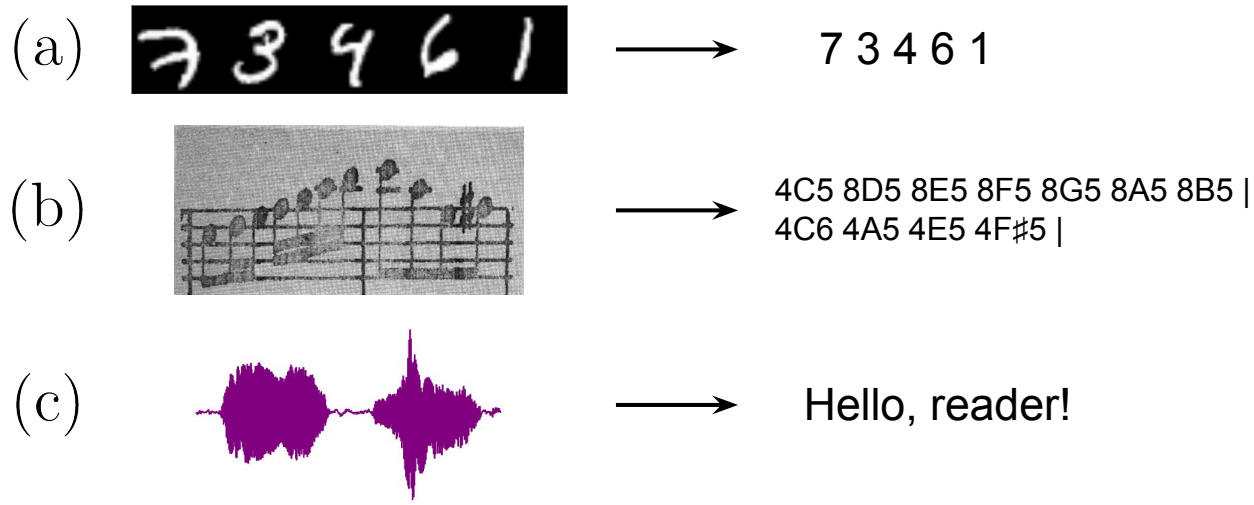


Figure 2.12: Examples of transcription tasks: (a) handwritten digits recognition (LeCun et al., 1989); (b) Optical Music Recognition (OMR) applied to the first music score scan published by Prerau (1972); and (c) a typical voice recognition scenario.

This includes notes/main melody estimation for singing voice (Rigaud and Radenen, 2016), instrument (Abesser and Müller, 2021) or over the global piece (Bittner et al., 2015). Others topics are chord identification (Jiang et al., 2019), drums transcription (Jacques and Roebel, 2019) or even several of these objectives at the same time (Ryynänen and Klapuri, 2008).

In the field of Optical Music Recognition (OMR), music scores are to be recognized inside images from high quality to highly deteriorated, such as pictures or scans of formatted scores (Calvo-Zaragoza et al., 2017) or handwritten ones (Baró et al., 2019).

Voice recognition

The voice community has heavily relied on recognition algorithms, and this thesis inherits from this research that can be traced back to the 1970s (Velichko and Zagoruyko, 1970) and remains today an active field of investigations (Liu et al., 2023). In the long quest to adapt daily experiences and knowledge on human-human interaction to human-machine communication, speech and singing recognition systems have played a crucial role in the development of, *e.g.*, vocal assistants (Michaely et al., 2017).

Automatic Speech Recognition (ASR) is the task dedicated to retrieve the *spoken* content in a *speech* recording. Automatic Lyrics Transcription (ALT) is the task dedicated to retrieve the *sung* content in a *singing* recording. As they are by essence extremely close, ASR algorithms serve as basis and are usually adapted to ALT. However, due to the more challenging nature of singing voice (cf. section 2.2.3), ALT performances are far below the ones reported on ASR.

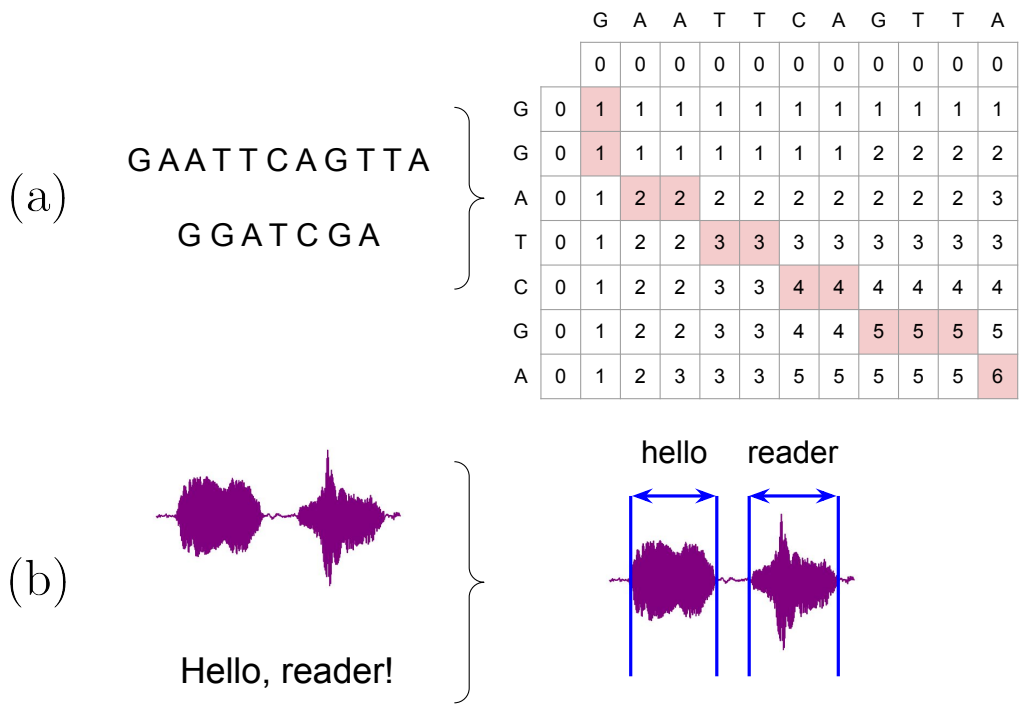


Figure 2.13: Examples of alignment tasks: (a) biological sequences alignment based on Needleman and Wunsch (1970)'s algorithm; and (b) a typical voice alignment scenario.

At the core of the scientific content defended in this thesis, the speech and singing recognition algorithms, especially relying on deep learning techniques, are explained in the next chapter.

2.3.3 Alignment algorithms

Similar to voice transcription, the sequence alignment algorithms are not unique to speech or singing processing. In fact, several research communities in the 1970s were facing the same need to find the optimal alignment path between sequences of several types and, as a result of their independent yet close investigations, similar algorithms were (re)discovered almost simultaneously. Some of the mentioned examples are shown in Figure 2.13.

General alignment

Telecommunications were the first field to benefit from a solution to sequence alignment when Viterbi (1967) introduced the fundamentals of his forward-backward decoding algorithm, which has become since then a core reference on all alignment systems and beyond (Ramasso et al., 2007). Many authors still refer to the "Viterbi's algorithm" or "Viterbi's decoding".

In bioinformatics, alignment systems have been widely used to arrange sequential data such as sequences of DNA, RNA, protein, amino acid, or equivalent through the work of Needleman and Wunsch (1970). The alignment of biological sequences have allowed to study their homology and identify regions of (dis)similarity between them. With such precise information, it has been possible to study gene expression (Aach and Church, 2001) and enhance shared structural patterns highlighting, *e.g.*, consequences of evolution, existence of common ancestor or effects of mutations (Ng and Henikoff, 2001).

In the broader scope of *time series* manipulation (Folgado et al., 2018), such alignments algorithms have been used to compare such series leading to human activity analysis (Machado et al., 2015) or on-line secure validation of handwritten signature (Xia et al., 2018).

Sequence alignments are also found in Natural Language Processing (NLP) with the example of bitext word alignment that consists in retrieving, given a text and its translation in another language, the translation relationships among the words (Brown et al., 1993).

Audio alignment

Alignment challenges were also studied in the audio literature. Synchronization and performance comparisons between musicians (or singers) can be achieved via audio-to-audio alignment (Kirchhoff and Lerch, 2011).

Audio-to-score alignment, which deals with mapping a music score and an associated audio recording, is also a well-known research problem (Simonetta et al., 2021), and can be used to analyse how a given performer (musician/singer) enacts a symbolic score. Its real-time variant, referred to as score following (Cont, 2009), can lead to on-the-fly page turning during a live performance (Arzt et al., 2008) or be the starting point towards adaptive and automatic accompaniment (Cont et al., 2012; Raphael, 2001).

Voice alignment

This thesis is primarily concerned with the alignment between audio recordings featuring voice and symbolic sequences. Automatic Speech Alignment (ASA) is the task dedicated to align a symbolic sequence with an associated *speech* recording. Automatic Lyrics Alignment (ALA) is the task dedicated to align a symbolic sequence with an associated *singing* recording.

ASA (resp. ALA) systems have been systematically built upon the recent advances on ASR (resp. ALT) – therefore, ASA and ALA are also based on similar models and ALA still faces more challenges than ASA because of the complexity of singing voice.

The scope of these voice-related paragraphs remain voluntarily limited to a simple introduction of voice recognition/alignment as the next chapter will extensively detail the technical background and design choices for their practical implementations – especially section 3.3 for legacy approaches and section 3.4 for deep learning-based strategies.

Section summary – Transcription and alignment tasks

Given a message m characterized by two emissions $\mathbf{x}$ and $\mathbf{y}$ with different modalities and/or expressivity,

- *Transcription* is the task aiming at predicting $\hat{\mathbf{y}}$ – an estimate of $\mathbf{y}$ – from $\mathbf{x}$;
- *Alignment* is the task aiming at mapping at least one dimension of each sequences $\mathbf{x}$ and $\mathbf{y}$ through $\hat{\boldsymbol{\pi}}^*$ – an estimate of the optimal path between them.

Sequence transcription and alignment problems arose in many literature and have been concerned with various objectives in audio and voice processing. They follow a similar two-stage architecture made of a modeling function $\mathcal{M}$ applied on $\mathbf{x}$ and a decoding module $\mathcal{D}$ specific to each task allowing the final predictions.

2.4 Applicative motivations of voice alignment

The central text-to-voice alignment problem, which historically emerged in the late 1970s, originated from a bottleneck in the speech recognition community: the need to automatically segment voice recordings into small labelled excerpts (*i.e.*, pairs of audio and their associated text), with the aim to build larger corpora than those available at that time. Ever since, the synchronization of a voice signal with its corresponding text has known many practical applications for the entertainment industry or the general public such as audio indexing via text, automatic subtitling or karaoke generation (Fujihara and Goto, 2012).

In this short section, the applicative motivations presented in this thesis are briefly mentioned to give a global overview on the launched initiatives and scientific and technological cooperations. The Chapter 7 will be dedicated to their practical accomplishment.

- *Concatenative phonetic synthesis*

  As mentioned, phonemes are the building blocks of voiced languages, from syllables to words and sentences etc. In the context of concatenative synthesis – a specific strategy for Text-to-Speech (TTS) – a target text is decomposed into its corresponding phonemes and audio is generated by selecting in an available dataset the most relevant voice samples associated with each phoneme and their surrounding context (usually, succession of two phonemes, *i.e.*, biphones, are used). Voice-to-phoneme alignment is prominent as the reference voice recordings needs to be aligned very precisely so that chosen regions for a given biphone make sense. The alignment algorithms of this thesis are applied to the alignment of new voices for ISiS[3]. See section 7.1 for further details.

[3]Ircam Singing Synthesis: https://forum.ircam.fr/projects/detail/isis/.

- *Production strategies of speech attitudes*

  Humans are clearly able to communicate social intentions when interacting with one another. In the context of such expressive speech, conveying noticeable social attitudes, the synchronization of audio with phoneme sequences allows group statistics between speakers regarding their speech prosody modulations and phonetic structure changes over time – a major step forward uncovering the interactional properties beyond speech prosody. See section 7.2 for further details.

- *Musicological analyses of singing voice style*

  Singers, very much like speakers, convey emotions and use deliberate effects to integrate expressivity in their performances. The study of singing style – *i.e.*, these production strategies at play in a singer's performances defining their very own artistic identity – is of great interest for musicologists eager to understand the artistic choices made by a singer. The automatic voice alignment algorithms developed in this thesis – *both* for syllables *and* notes – fully inscribe themselves in such musicological research as they highly simplify tedious tasks traditionally done manually and allow exploring in details, *e.g.*, rythmic, articulation, or multi-modal gesture like melisma. To this aim, a musicological pipeline for singing voice style analysis based on neural voice processing and alignment has been developed in the **ARS** project and benefit from this thesis research. See section 7.3 for further details.

- *Perspectives*

  Finally, some ongoing/future applications are mentioned in section 7.4, two of which are

  – Automatic pre-segmentation of large corpora like turning an audiobook into small coherent pairs of audio and text, which is possible thanks to a collaborative work dedicated to a linear memory decoding algorithm allowing to alignm very long audio recordings, *e.g.*, full audiobooks or lengthy music playlists and their transcripts;

  – Joint singing voice separation and alignment so that both tasks can help each other as suggested by recent literature.

Section summary – Applicative motivations of voice alignment

Several applications of voice alignment motivates the research in this thesis. Phoneme-level alignment paves the way for singing concatenative synthesis and understanding the production strategies involved in speech attitudes. Word-level alignment, which is the standard case study, is also of high interest, especially when applied on very long audio recordings. Coupled with note alignment, it greatly simplifies tedious and time-consuming processes required in musicological analyses.

2.5 Context and applicative motivations in a nutshell

Chapter summary – Context and applicative motivations

In this chapter, I exposed the key concepts to carry out voice analysis research, with a focus on voice alignment and its applications. Based on human communication theory, I introduced the two modalities at the heart of my research: *symbolic sequences* and *voice*. Sequences inherit from a long history of writing languages and require the existence of an alphabet of characters $\mathcal{A}$ – like the Latin alphabet $\mathcal{A}_\ell$ or note alphabet $\mathcal{A}_\text{♩}$. From a signal processing perspective, I presented two relevant representations to analyse and synthesize voice: (1) the *log-mel-spectrogram* – capturing all perceptual features; and (2) the *spectral envelope* – characterizing well formants and phonetic content. Finally, I defined the closely related *transcription* and *alignment* tasks and mentioned the thrilling applications offered by synchronizing voice with symbols, including the ones this thesis is concerned with, *e.g.*, musicological studies, production of vocal attitudes, notes alignment, etc.

Chapter 3

Scientific background for voice alignment

*"Nothing in life is to be feared, it is only to be understood. Now is the **Time** to understand more, so that we may fear less."*

– Marie Skłodowska-Curie

This chapter aims to provide a comprehensive review of the mathematical concepts, technical tools and scientific strategies for designing voice-to-symbol synchronization systems and the practical implementations of their associated acoustic models and decoding modules.

Upon delving into the essentials of *deep learning* in section 3.1, which form the foundation for the rest of this thesis, and providing a formalization of the *voice alignment problem* in section 3.2, the various options found in the literature for the acoustic modeling and the decoding module are presented. To this aim, the section 3.3 is dedicated to the *traditional approaches* and notably introduces the Dynamic Time Warping (DTW) algorithm as a reference decoding module and retraces the emergence of probabilistic acoustic models. Following recent and promising trends, the section section 3.4 focuses on the latest *deep learning techniques* applied to voice alignment with an exploration in greater depth, in section 3.5, of the *Connectionist Temporal Classification (CTC)* which, given its flexibility, has been chosen as the main strategy to perform voice-to-symbol alignment in this thesis. This chapter is summarized in section 3.6.

3.1 Deep learning background

The emergence of deep learning approaches, and especially deep neural networks, has initiated a breakthrough in many scientific fields. For a given task at hand, a deep learning-based model can specialize "itself" by automatically targeting and exploiting the most relevant task-related information in the data, thus outperforming legacy methods that were rather dependent on fixed prior knowledge. Speech and singing voice communities are no exception as recent analysis and synthesis systems have been built upon deep learning techniques, increasingly replacing former approaches, while denoting major progress in comparison to previous literature.

This section reviews the motivations and the mathematical notions necessary to carry out research using deep learning as a tool for voice processing tasks (*i.e.*, voice transcription and alignment). A brief contextualization on data-driven systems serves as basis to further introduce general machine/deep learning concepts, notably deep neural networks and their architecture design. For an extensive explanation of deep learning, one can refer to (Goodfellow et al., 2016), the pointing reference of this section.

3.1.1 Data-driven approach

Deep learning is the computer science field dedicated to the development of algorithms capable to *learn* specific and relevant patterns from data according to some objective. These algorithms are known as as *data-driven* systems as they exploit the observation of a wide quantity of data.

In opposition to *classic* systems − in the denomination of Humphrey et al. (2012) − deep learning models do not require expert prior knowledge for defining *hand-crafted* 1) features and/or 2) prediction rules from them. Instead, and from the data themselves, such a system can manipulate complex abstractions, automatically extract relevant features and make predictions. Frequently, instead of the raw data, compact and interpretable hand-crafted representations, *e.g.*, Time-Frequency Representations (TFR) for audio, are computed at the very beginning of the system − but are eventually turned into learned features by the model.

Benefiting from greater computational power and the latest improvements in hardware and software infrastructures, deep learning-based systems have shown impressive generalization properties that have lead to significant advances in many states of the art, *e.g.*, computer vision (Loiseau et al., 2021; Voulodimos et al., 2018), natural language processing (Otter et al., 2020) and audio domain (Briot, 2022; Douwes et al., 2023; Purwins et al., 2019).

3.1.2 Essentials

In this thesis, feature extraction and task-oriented prediction rely on the training of Deep Neural Networks (DNN) in a supervised learning scenario, obtaining their parameters by optimization of a loss function via mini-batch gradient descent. These core concepts are gradually developed.

Artificial neuron

A whole category of learning models have been vaguely inspired from the human brain and specifically from a simplified version of its functioning centered around the notion of *neuron*. Let $\mathscr{X}^*$ and $\mathscr{W}$ denote some feature space definition ensemble with no *a priori*.

In imitation of a *biological* neuron, an *artifical* neuron receives an input excitation $\mathbf{x} \in \mathscr{X}^*$ to which it is more or less sensitive. The reactivity of the neuron to the excitation is embodied by weighting parameters $\mathbf{w} \in \mathscr{W}$ and a response function $f : \mathscr{W} \times \mathscr{X}^* \to \mathbb{R}$ that applies some transformation to $\mathbf{x}$ according to $\mathbf{w}$. By electrical analogy, if the resulting "potential" $f(\mathbf{w}, \mathbf{x}) \in \mathbb{R}$ is sufficient, *e.g.*, higher than a threshold $b \in \mathbb{R}$, then the neuron is said to be activated and can transmit information to the next neuron, and so on. This neural activation, which is intrinsically non-linear, is simulated by an *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$.

Summing up, an artificial neuron predicts an output $\hat{y} \in \mathbb{R}$ from inputs $\mathbf{x} \in \mathscr{X}^*$ following

$$\hat{y} \;=\; \sigma\left(f(\mathbf{w}, \mathbf{x}) + b\right) \tag{3.1}$$

and weights $\mathbf{w}$ and bias b are referred to as parameters of the neuron.

Non-linear activation

The non-linearity introduced in a neural response is of utmost importance in deep learning as it is the core reason a neuron can represent (and learn) complex relationships between its outputs and inputs.

There is a wide range of options for choosing the activation function σ, three of which will be used in this manuscript:

- The Rectified Linear Unit (ReLU) introduced by Maas et al. (2013) and defined $\forall x \in \mathbb{R}$ as
$$\mathrm{ReLU}(x) \;=\; \max(0, x) \tag{3.2}$$
- The hyperbolic tangent, smoother than ReLU and with a bounded output range $[-1, 1]$, such that $\forall x \in \mathbb{R}$,
$$\tanh(x) \;=\; \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.3}$$
- The softmax function, which returns a probability distribution (hence a bounded output range $[0, 1]$) from the inputs it receives, *i.e.*, $\forall \mathbf{x} \in \mathbb{R}^T$, $\mathrm{softmax}(\mathbf{x}) \in \mathbb{R}^T$ with
$$\mathrm{softmax}(\mathbf{x})_t \;=\; \frac{\mathbf{x}_t}{\sum_{t=0}^{T-1} \mathbf{x}_t} \tag{3.4}$$
where the $t \in \{0, \ldots, T-1\}$ subscript denotes the tth element of its associated vector.

Deep Neural Networks (DNN)

Eq. (3.1) represents the response of a neuron to an excitation. In practice, there is hardly ever a single neuron but rather a *layer* of neurons simultaneously excited. For a given layer ℓ composed of $n \in \mathbb{N}$ neurons, the previous formalism remains identical when considering the weights of the entire layer $\mathbf{W}_\ell \in \mathscr{W}^n$ as the ensemble of all neurons' weights, their associated response function $\boldsymbol{f}$, and the outputs $\hat{\mathbf{y}}_\ell \in \mathbb{R}^n$ and bias $\mathbf{b}_\ell \in \mathbb{R}^n$ as vectors instead of scalars:

$$\hat{\mathbf{y}}_\ell \;=\; \boldsymbol{f}_\ell(\mathbf{x}) \;=\; \sigma\left(\boldsymbol{f}(\mathbf{W}_\ell, \mathbf{x}) + \mathbf{b}_\ell\right). \tag{3.5}$$

The weights $\mathbf{W}_\ell$ and bias $\mathbf{b}_\ell$ are the parameters of the neural layer and can be changed to orientate the outputs $\hat{\mathbf{y}}_\ell$ towards certain values, as it will be further detailed.

In the human brain, layers of neurons are inter-connected to one another, entangled in a intricate *network*, and propagate information step by step up to specialized regions of the brain. Artificial Neural Networks (ANN) follow the same approach as they are defined by a succession of layers, each composed of several neurons. Deep Neural Networks (DNN) are artificial networks sufficiently *deep*, that is composed of "many" layers − although there is no concrete consensus on how much "many" should be.

Layer after layer, more and more parameters and nonlinearities allow increasingly complex modeling of the data $\mathbf{x}$ towards the final prediction $\hat{\mathbf{y}}$. For $\ell = 0, \ldots, l-1$ indexing the layers, the prediction of the DNN, which is the output last layer, is expressed as:

$$\hat{\mathbf{y}} \;=\; \underbrace{(\boldsymbol{f}_{l-1} \circ \ldots \circ \boldsymbol{f}_\ell \circ \ldots \circ \boldsymbol{f}_0)}_{=\boldsymbol{f}_\Theta}(\mathbf{x}) \;=\; \boldsymbol{f}_\Theta(\mathbf{x}). \tag{3.6}$$

where Θ refer to all trainable parameters $\{\mathbf{W}_\ell, \mathbf{b}_\ell\}_{\ell \in [0, l-1]}$ − often simply called *weights* through misuse of language. The neural prediction can then be supervised, *i.e.*, compared to a reference.

Supervised learning

Supervised learning is a paradigm in which data are annotated and these ground truths are used as references to evaluate and update the weights and bias of deep neural networks. Given a dataset $\mathscr{D}$ of labeled input-output pair $(\mathbf{x}, \mathbf{y})$, the label $\mathbf{y}$ associated with representation $\mathbf{x}$ is what is expected to be predicted by the final output $\hat{\mathbf{y}} = \boldsymbol{f}_\Theta(\mathbf{x})$ of a neural model.

It is desired that $\boldsymbol{f}_\Theta$ perfectly fits an unknown function $\boldsymbol{f}^*$ that fully reports on the connection between the labels $\mathbf{y}$ and the representations $\mathbf{x}$, *i.e.*, $\boldsymbol{f}^*(\mathbf{x}) = \mathbf{y}$. This is barely possible in practice as the design choices for $\boldsymbol{f}_\Theta$ typically induce *inductive bias* preventing from *exactly* matching $\boldsymbol{f}^*$. The *universal approximation theorem*, though, states that any neural network properly configured can represent a whole family of functions with only few hypotheses on the activation function of each neuron (verified by, *e.g.*, ReLU, tanh and softmax). Deep neural networks are therefore good candidates for *approximating* via $\boldsymbol{f}_\Theta$ the underlying function $\boldsymbol{f}^*$.

This research is exclusively based on supervised learning but systems can also rely on *semi-supervised* learning (Zhu, 2005) when only some part of the data are annotated or even *unsupervised* learning (Barlow, 1989) when no annotation is accessible. Supervised learning is heavily dependent on the quality of the annotations as they are, by nature, considered as references to predict. Wrongly labeled pair can therefore induce severe errors. Annotation checking is in itself an activate research field in deep learning (Meseguer-Brocal et al., 2020a; Ridzuan and Zainon, 2019). In this work, however, all annotations are considered correct even if manual inspections or automatic checking of the labels are not systemically done.

Loss function

In order to determine to which extent the model prediction $\boldsymbol{f}_\Theta$ is a relevant approximation of $\boldsymbol{f}^*$, and tune its parameters towards a better fit if need be, a quantitative measure of the "proximity" between $\boldsymbol{f}^*$ and $\boldsymbol{f}_\Theta$ is essential.

This is done by defining a *loss function* $\mathscr{L}(\Theta)$ – or *cost function*, *objective function* or *error function* – which compares the neural prediction to the ground truth for each pair $(\mathbf{x}, \mathbf{y})$ contained in the dataset $\mathscr{D}$ given some distance $\mathbb{d}$ (not necessarily a *distance* in the mathematical sense). It reads:

$$\mathscr{L}(\Theta) \;=\; \sum_{(\mathbf{x},\mathbf{y})\in\mathscr{D}} \mathbb{d}\left[\hat{\mathbf{y}}||\mathbf{y}\right] \;=\; \sum_{(\mathbf{x},\mathbf{y})\in\mathscr{D}} \mathbb{d}\left[\boldsymbol{f}_\Theta(\mathbf{x})||\mathbf{y}\right]. \tag{3.7}$$

The choice of the cost/distance $\mathbb{d}$ depends on the task at hand. Several distances will be manipulated in this manuscript, as a core part of this research is precisely dedicated to emphasize a correct combination of cost functions. These will be introduced whenever necessary.

Optimization

The goal of a deep learning algorithm is to minimize the value of its associated loss function $\mathscr{L}(\Theta)$ as the smaller the loss, the better $\boldsymbol{f}_\Theta$ fits the true function $\boldsymbol{f}^*$ to uncover. Concretely, one is looking for the theoretical optimal parameters Θ^* such that

$$\Theta^* \;=\; \underset{\Theta}{\operatorname{argmin}}\, \mathscr{L}(\Theta). \tag{3.8}$$

This is an *optimization* problem which can be solved thanks to numerical methods allowing to progressively update the parameters Θ while diminishing the loss function – with the aim to eventually reach sufficiently adequate weights and bias (*i.e.*, finding a *local* minimum close enough to the *global* minimum).

Back-propagation & gradient descent

Back-propagation algorithms (Rumelhart et al., 1986) are commonly used to perform the optimization of a model. Their philosophy is to "learn from made mistakes".

To do so, given the current weights and bias Θ, the loss value $\mathscr{L}(\Theta)$ is computed. Then, its partial derivate or *gradient* with respect to the model's parameters is calculated, *i.e.*,

$$\nabla_\Theta \mathscr{L}(\Theta) \quad = \quad \frac{\partial \mathscr{L}}{\partial \Theta}(\Theta). \tag{3.9}$$

As a positive gradient indicates the direction augmenting the error for parameters Θ, the opposite of the gradient is the direction to follow for decreasing the loss.

The deep learning community has extensively relied on *gradient descent* to update step by step the parameters Θ in this direction. The kth modification of the weights and bias is

$$\Theta_k \quad \leftarrow \quad \Theta_{k-1} - \lambda \nabla_\Theta \mathscr{L}(\Theta_{k-1}). \tag{3.10}$$

This is known as a *training step* in which λ − the *learning rate* − regulates the strength of the update. *Training* deep neural networks consists in successively applying this procedure.

More sophisticated and adaptive gradient descent algorithms have been proposed in order to both speed up convergence and avoid local minima. An efficient approach is to not only consider the gradient, but also its momentums (*e.g.*, mean and variance) in the update rule. This is the strategy employed by the ADAM optimizer (Kingma and Ba, 2014). Ubiquitous in the literature, it will be systemically used for training in this work.

Minibatch gradient descent

In practice, the datasets $\mathscr{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_i$ are too large to allow computation of the loss value as expected in Eq. (3.7), given the limited resources and available memory on modern machines.

An alternative is to rely on *minibatch* gradient descent to compute an approximate of the loss and its gradient. Instead of considering the whole database $\mathscr{D}$, a few numbers of samples (*i.e.*, a *minibatch*) are randomly selected and the loss function is calculated only on this subset of examples. The gradient can be computed and the weights and bias updated following above-mentioned procedure. The size of the minibatch − the *batch size* − is a key hyperparameter for training a model, very much like the learning rate.

Model generalization

Measuring the loss value during the training procedure does not indicate the *generalization* properties of the model, *i.e.*, its capacity to predict coherent results when applied to new data outside of the dataset $\mathscr{D}$.

Indeed, the network may become too specialized on the data seen during training and unable to deal with data outside of this scope. In order to prevent such an *over-fitting*, the dataset $\mathscr{D}$ is generally split into two subsets referred to as the *training set* and the *validation set*.

The training set is used to update the model parameters via minibatch gradient descent. The validation set, containing different data, ensures that the model still generalizes by monitoring that the validation loss (or a validation metric) keeps decreasing alongside the training loss. The training procedure can be interrupted as soon as over-fitting is dedected: this is known as *early stopping*.

Another prevalent option to counter the appearance of over-fitting is *dropout* (Srivastava et al., 2014), which consists in randomly deactivating several neurons between two layers ℓ and $\ell + 1$. This simple strategy allows independent and efficient joint learning of the many sub-network composing the whole network.

3.1.3 Architecture design

Previous section presented Deep Neural Networks (DNN) and their training to optimize the parameters Θ defining an approximative, yet robust, predictive function $\boldsymbol{f}_\Theta$. According to Eq. (3.6), this learned function can be decomposed into many successive layers of neurons, $\boldsymbol{f}_\ell$ with $\ell \in \{0, \dots, l-1\}$. Among the available options to design the architecture of a neural layer, each associated with a different $\boldsymbol{f}_\ell$, the ones relevant to this thesis are exposed.

Dense

A *dense* layer − or fully connected layer − connects all neurons to the input representations with a simple multiplicative weighting:

$$\hat{\mathbf{y}}_\ell \;=\; \boldsymbol{f}_\ell(\mathbf{x}) \;=\; \sigma\left(\mathbf{W}_\ell \mathbf{x} + \mathbf{b}_\ell\right). \tag{3.11}$$

Dense layers are the building block of Feed-Forward Neural Networks and remain often used as final layer in many practical cases.

Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) (LeCun et al., 1995) exploit convolution operation − as presented in Eq. (2.12) or Eq. (2.13) according to data dimension −, to extract abstractions from filtering the data. The learnable weights $\mathbf{W}_\ell$ constitute the convolution kernel, such that

$$\hat{\mathbf{y}}_\ell \;=\; \boldsymbol{f}_\ell(\mathbf{x}) \;=\; \sigma\left(\mathbf{W}_\ell * \mathbf{x} + \mathbf{b}_\ell\right). \tag{3.12}$$

Frequently, a single convolutional layer ℓ is *multi-channel*, meaning that multiple kernels are learned and applied to the same input $\mathbf{x}$, thus resulting in several filtering of the data.

Due to the nature of the convolution operation, the dimensions on which the kernel(s) is(are) slid across are susceptible to change. *Padding* of the input data can be done before the convolution to keep the(se) dimension(s) identical. This is systematically done in this thesis.

Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNN) (Rumelhart et al., 1986) are specifically designed to process sequential data by introducing internal memory and dependence between components of the prediction. Denoting $\mathbf{x} = \{\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(T-1)}\}$ and $\hat{\mathbf{y}}_\ell = \{\hat{\mathbf{y}}_\ell^{(0)}, \dots, \hat{\mathbf{y}}_\ell^{(t)}, \dots, \hat{\mathbf{y}}_\ell^{(T-1)}\}$, a typical recurrent layer computes:

$$\forall t \in \{0, \dots, T-1\}, \quad \hat{\mathbf{y}}_\ell^{(t)} \;=\; \sigma\left(\mathbf{W}_\ell^y \hat{\mathbf{y}}_\ell^{(t-1)} + \mathbf{W}_\ell^x \mathbf{x}^{(t)} + \mathbf{b}_\ell\right). \tag{3.13}$$

The learnable weights are thus separated into recurrent weights processing both current input vector $\mathbf{W}_\ell^x$ and previous internal output state $\mathbf{W}_\ell^y$. If causaliy is not a key requirement for the foreseen application, one can rely on *bidirectional* recurrent layer (Schuster and Paliwal, 1997) by considering Eq. (3.13) for the reserved input sequences $\mathbf{x}_r = \{\mathbf{x}^{(T-1)}, \dots, \mathbf{x}^{(0)}\}$ and concatenate causal and anti-causal predictions.

Well-known drawbacks of basic RNNs are the vanishing and exploding gradients problems induced by their recurrences (Bengio et al., 1993). As a result, RNN cells with internal memory reset were introduced to specifically address these limitations. The Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) is one of the widely used option.

Attention mechanism

Attention mechanisms allow to learn the relative importance between the components of two sequences and integrate this information into a third sequence. They were introduced by Bahdanau et al. (2014) for the machine translation task, a problem for which the structure between input and output sequences may differ and a general context (*e.g.*, at the sentence level) can be highly beneficial to learn the structural changes.

A Bahdanau attention (also known as *additive attention*) layer, therefore, learns a weighting between a *query* vector $\mathbf{q} \in \mathbb{R}^{T \times E}$ and a *key* vector $\mathbf{k} \in \mathbb{R}^{M \times E}$ and generates an attention context from a value vector $\mathbf{v} \in \mathbb{R}^{M \times E'}$. It reads, with T the transpose operator,

$$\hat{\mathbf{y}}_\ell \;=\; \boldsymbol{f}_\ell(\mathbf{k}, \mathbf{q}, \mathbf{v}) \;=\; \text{softmax}\left(\mathbf{q}\mathbf{W}_\ell \mathbf{k}^{\mathrm{T}} + \mathbf{b}_\ell\right)\mathbf{v}. \tag{3.14}$$

Note that the query and the key must share their last dimension E and that the key and the value must share their first dimension M. In practice, the key and the value are often equal $\mathbf{k} = \mathbf{v}$. Self-attentions are also considered when $\mathbf{k} = \mathbf{q} = \mathbf{v}$. Architectures relying partially – or even solely – on (self-)attention mechanisms, *e.g.*, Transformer (Vaswani et al., 2017), have become popular design choices for DNN.

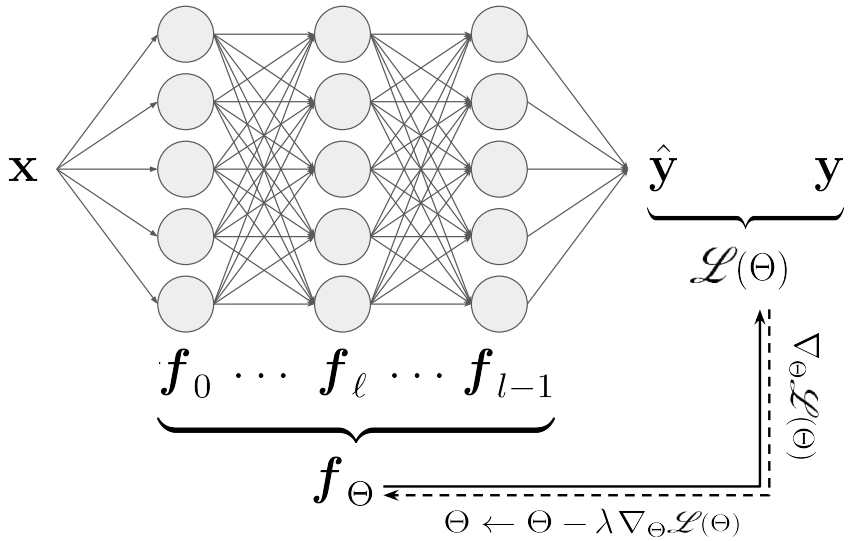


Figure 3.1: A generic deep neural network. Inputs $\mathbf{x}$ are turned into predictions $\hat{\mathbf{y}}$ thanks to the modeling function $\boldsymbol{f}_\Theta$ whose parameters Θ are updated via gradient descent on the loss function $\mathscr{L}$ measuring the proximity between $\hat{\mathbf{y}}$ and ground truth $\mathbf{y}$.

Batch Normalization (BN)

Batch Normalization (BN) is a method proposed by Ioffe and Szegedy (2015) to fasten the training of Artificial Neural Networks (ANN) and make this procedure more stable. It consists in normalizing the inputs received by a layer by re-centering and re-scaling of its values.

Section summary – Deep learning background

With the emergence of large annotated datasets $\mathscr{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_i$ and progress in computer infrastructures, Deep Neural Networks (DNN) were shown to be worthy candidates to estimate the underlying function describing the relationships between representation-label pairs $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_i$ with a non-linear, Θ-parameterized function $\boldsymbol{f}_\Theta$. This function can be implemented via a succession of convolutional, recurrent, attention and/or dense layers connected with non-linear activations. The weights and bias Θ of the model are typically obtained through the progressive optimization, *i.e.*, minimization, of a loss function $\mathscr{L}(\Theta)$ by minibatch gradient descent under ADAM update rule. The Figure 3.1 proposes a visual summary of all of these notions.

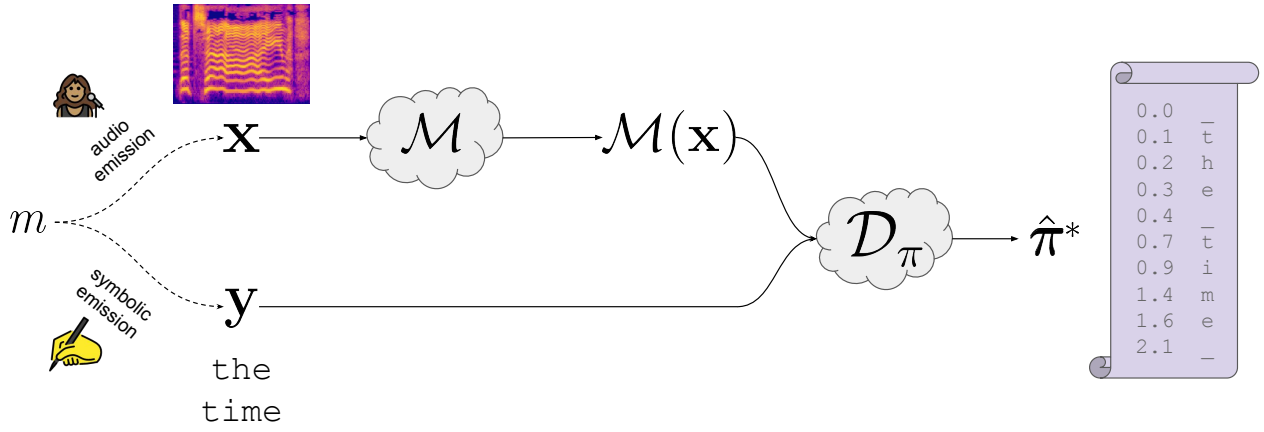


Figure 3.2: The voice alignment problem – finding the best time-symbol correspondence between two message emissions from audio (log-mel-spectrogram) and symbolic (texts, notes) domains.

3.2 Voice alignment problem statement

In this short section, the voice alignment problem is clearly stated, with an emphasis on the preparation of the voice data.

3.2.1 Formalization

The Figure 3.2 depicts how the generic sequence-to-sequence alignment can be applied to the specific voice alignment problem.

In that case, data to synchronize represent an audio recording featuring voice with a *temporal dimension* of interest, and a symbolic sequence of characters (text alignment) or notes (melody alignment). The goal, in aligning these sequences, is to map each label (or symbol) to an *instant* and a *duration* or, which is equivalent, a *start and end time markers*.

The modeling function $\mathcal{M}$ aims to generate a *timestamped encoding* of the audio, *i.e.*, it creates a hidden representation which conserves a temporal axis, essential for temporal alignment. As this model processes audio data only and aims to extract relevant voice-related parameters, it is often referred to as *acoustic model* by the speech community (Hinton et al., 2012). The resulting latent code, or timestamped encoding $\mathcal{M}(\mathbf{x})$, usually goes by the name of *posteriorgram* or *saliency map* so that notation $\mathcal{P} \equiv \mathcal{M}(\mathbf{x})$ will be chosen for such hidden encoding.

The decoding module, $\mathcal{D}_\pi$ is generally based on Dynamic Time Warping (DTW) or equivalent. This work, however, will rely on a specific decoding module as it uses a specific framework, known as Connectionist Temporal Classification (CTC) presented in a later section.

3.2.2 Voice data

Voice representations

In this work, the inputs $\mathbf{x}$ of all neural network are 2-D Time-Frequency Representations (TFR) such as (log-mel-) spectrograms or spectral envelope as introduced in section 2.2. They are accounting for the oral modality of a message communicated through voice (see section 2.1.3).

Their *temporal* axis is indexed by $t \in \{0, \dots, T-1\}$ where T is the number of computed frames. Their *frequency* axis is indexed by $f \in \{0, \dots, F-1\}$ where F is the number of bins.

Symbolic representations

The supervised labels $\mathbf{y}$ a neural network aims to predict are sequences of symbols belonging to an alphabet $\mathcal{A}$ of size L. In order to manipulate and process sequences of numbers, each element of $\mathcal{A}$ is mapped to a unique figure of the ensemble $\mathcal{A}' = \{0, \dots, L-1\}$.

An one-hot encoding of the labels is also possible to turn them into a 2-D representation: for a symbol $s \in \mathcal{A}'$ a vector $[\delta_s^0, \dots, \delta_s^{L-1}]$ is generated with δ_s^k being the KRONECKER delta returning 1 when $s = k$ and 0 otherwise, and such vectors can be concatenated for each label.

The following examples illustrates these mappings, *i.e.*, the sequence-to-number conversion and the final one-hot encoding:

$$\mathbf{y} \;=\; \text{b a ø a b} \;\equiv\; 1\ 0\ L\text{-}1\ 0\ 1 \;\equiv\; \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The labels account for the symbolic modalities of a message communicated in a voice signal. As defined in section 2.1.2, the Latin alphabet $\mathcal{A}_\ell$ will be used for characters and the note alphabet $\mathcal{A}_\text{♩}$ for melodic lines.

Their *sequential* axis is indexed by $m \in \{0, \dots, M-1\}$ where M is the length of the label sequence. Their *encoding* axis, if present, is indexed by $l \in \{0, \dots, L-1\}$ with L the alphabet size.

Section summary – Voice alignment problem statement

Voice alignment systems are based on a two-step principle made of:

- an *acoustic model* $\mathcal{M}$ generating a timestamped encoding $\mathcal{P} \equiv \mathcal{M}(\mathbf{x})$ of the audio voice recordings $\mathbf{x}$;
- and a *decoding module* $\mathcal{D}_\pi$ mapping the timing information from $\mathcal{P}$ to the target symbols to synchronize $\mathbf{y}$.

Concretely, inputs $\mathbf{x} \in \mathbb{R}^{T\times F}$ are Time-Frequency Representations (TFR) and labels $\mathbf{y} \in \{0,\dots,L-1\}^{M\times 1}$ or $\{0,1\}^{M\times L}$ are symbolic sequences such as texts (lyrics, syllables, phonemes) or notes according to an alphabet $\mathcal{A}$ of size L.

3.3 Traditional approaches to voice alignment

Over the years, the forced alignment problem has been implemented with multiple strategies. In this section, an historical literature review presents the traditional approaches that have been used to tackle voice alignment. The decoding module $\mathcal{D}_\pi$ involved in an alignment system has been quite systematically based on Dynamic Time Warping (DTW) inheriting from sequence-to-sequence similarity research. The acoustic modeling $\mathcal{M}$, however, has been built upon several strategies from hand-crafted systems to the noteworthy emergence of stochastic approaches and notably Hidden Markov Models (HMM).

3.3.1 Dynamic Time Warping (DTW)

A naive determination of the alignment, as defined by the optimal path from Eq. (2.23) and illustrated in Figure 2.9, based on path exploration is usually intractable as there are too many possible paths between the two sequences $\mathbf{x} \in \mathscr{X}^*$ and $\mathbf{y} \in \mathscr{Y}^*$. Indeed, as reported by Garreau et al. (2014), the set $\Pi(\mathscr{X}^*, \mathscr{Y}^*)$ contains all paths on a rectangular grid starting from the northwest $(0,0)$ and ending at the southeast $(T-1, M-1)$ corners. Its cardinality defines the Delannoy numbers (Banderier and Schwer, 2005). In the case of an infinitely large square grid (*i.e.*, $T = M$ and $T \to +\infty$), Torres et al. (2003) showed that:

$$\#\Pi(\mathscr{X}^*, \mathscr{Y}^*) \simeq \frac{\left(3+2\sqrt{2}\right)^T}{\sqrt{\pi T}\sqrt{3\sqrt{2}-4}}. \tag{3.15}$$

Despite being finite values, examples for $T = 50$ (10^{37} paths) and for $T = 100$ (10^{75} paths) clearly points out the impossibility to calculate each path independently and keep the best one.

In the early 1970s, researchers from various communities (Needleman and Wunsch, 1970; Velichko and Zagoruyko, 1970; Vintsyuk, 1968; Viterbi, 1967) needed to solve the sequence alignment problem but were confronted to this specific issue.

Interestingly, they all proposed a similar algorithm noticing that the best path could be computed indirectly via Dynamic Programming (DP) techniques. As a result, this algorithm was named Dynamic Time Warping (DTW) by speech specialists (Vintsyuk, 1968).

The core principle of DTW is that the optimal similarity measure between two sequences can actually be obtained recursively, considering the sequence prefixes at intermediate indexes t and m, that is $\mathbf{x}_{0:t} = \{\mathbf{x}_0, \dots, \mathbf{x}_t\}$ and $\mathbf{y}_{0:m} = \{\mathbf{y}_0, \dots, \mathbf{y}_m\}$, respectively.

To this aim, an accumulative score matrix $\boldsymbol{\alpha} \in \mathbb{R}^{T \times M}$ is defined so that $\boldsymbol{\alpha}[t, m]$ represents the optimal alignment similarity between the prefixes $\mathbf{x}_{0:t}$ and $\mathbf{y}_{0:m}$. The computation of $\boldsymbol{\alpha}$ involves the similarity measure $\mathscr{S}$ from Eq. (2.23), and the set of permitted gaps $\mathcal{G}$ between successive states introduced in Eq. (2.22).

The classical *textbook* DTW algorithm (Itakura, 1975) allows insertions, deletions and substitutions between the two sequences such that $\mathcal{G} = \{(1,1), (0,1), (1,0)\}$. The initialisation steps and recursion rules for matrix $\boldsymbol{\alpha}$ are therefore:

$$\begin{aligned}
t = 0 \quad \forall m \quad \boldsymbol{\alpha}[0, m] &= \sum_{k=0}^{m} \mathscr{S}(\mathbf{x}_0, \mathbf{y}_k) \\
\forall t \quad m = 0 \quad \boldsymbol{\alpha}[t, 0] &= \sum_{k=0}^{t} \mathscr{S}(\mathbf{x}_k, \mathbf{y}_0) \\
\forall t > 0 \quad \forall m > 0 \quad \boldsymbol{\alpha}[t, m] &= \mathscr{S}(\mathbf{x}_t, \mathbf{y}_m) + \max \begin{cases} \boldsymbol{\alpha}[t, m-1] \\ \boldsymbol{\alpha}[t-1, m] \\ \boldsymbol{\alpha}[t-1, m-1]. \end{cases}
\end{aligned} \tag{3.16}$$

In the context of temporal alignment (*i.e.*, t represents a time quantity), the matching between the sequences may involve a strict temporal causality such that $\boldsymbol{\alpha}[t, m]$ cannot depend on $\boldsymbol{\alpha}[t, m'], \forall m' \in \{1, \dots, m-1\}$. In that case, the permitted transitions are reduced to $\mathcal{G} = \{(1,0), (1,1)\}$. It comes:

$$\begin{aligned}
t = 0 \quad \forall m \quad \boldsymbol{\alpha}[0, m] &= \sum_{k=0}^{m} \mathscr{S}(\mathbf{x}_0, \mathbf{y}_k) \\
\forall t \quad m = 0 \quad \boldsymbol{\alpha}[t, 0] &= \sum_{k=0}^{t} \mathscr{S}(\mathbf{x}_k, \mathbf{y}_0) \\
\forall t > 0 \quad \forall m > 0 \quad \boldsymbol{\alpha}[t, m] &= \mathscr{S}(\mathbf{x}_t, \mathbf{y}_m) + \max \begin{cases} \boldsymbol{\alpha}[t-1, m] \\ \boldsymbol{\alpha}[t-1, m-1]. \end{cases}
\end{aligned} \tag{3.17}$$

At each iteration, local scores are accumulated and only the best prefix of the optimal path is kept − this is known as the *forward pass*.

The global optimal path $\hat{\boldsymbol{\pi}}^*$ between the sequences can be retrieved by backtracking the retained transitions that occurred during the forward pass − this is known as the *backward pass*. In the end, once all $\boldsymbol{\alpha}$s have been computed, the optimal alignment similarity is simply:

$$\mathbf{A}^* = \boldsymbol{\alpha}[T-1, M-1]. \tag{3.18}$$

The DTW algorithm has found numerous applications in Music Information Retrieval (MIR) tasks (Müller, 2007a) and audio alignment including melody search (Mongeau and Sankoff, 1990), music tracking (Arzt, 2016), score following (Dixon, 2005), audio matching (Müller et al., 2005), beat tracking (Ellis, 2007) or version identification (Serra and Gómez, 2007). In practice, the decoding module $\mathcal{D}_\pi$ of a forced alignment system generally relies on DTW or one of its variant.

3.3.2 Pioneer acoustic modeling

In the first stages of voice alignment research, in the 1980s, pioneering works were exploiting hand-crafted strategies in order to force-align sequences to audio.

The synchronization of phonetic utterances to audio typically relied on (1) features extracted from the audio thanks to signal processing techniques; and (2) domain knowledge rules for the forced alignment step *per se* (Leung and Zue, 1984; Wagner, 1981).

Similarly, in the early appearances of score following in 1984 (Dannenberg, 1984; Vercoe, 1984), the alignment challenge was actually dealt with string matching techniques in real-time instead of direct use of the audio stream or spectrograms, due to the limited computational capabilities at that time.

3.3.3 Hidden Markov Models (HMM)

A key noteworthy progress in Automatic Speech Recognition (ASR) and, consequently, audio alignment literature was the emergence of approaches based on probability theory during the 1990s (Grubb and Dannenberg, 1994; Ljolje and Riley, 1991; Placeway and Lafferty, 1996). The main motivation for these was the growing necessity to handle all different kinds of uncertainties or unpredictable events in audio performances such as mispronunciation or mistakes or skips, *e.g.*, (Gupta et al., 2017; Nakamura et al., 2015).

Probabilistic approaches, therefore, tend to have better applicative flexibility than previous pioneering works. Most of them have been based on *generative* probabilistic models whose main assumption is that the audio recording has been *generated* by the events reported in the sequences to align, *e.g.*, the music score or the list of phonemes.

Maybe the major breakthrough in alignment research was the application of Hidden Markov Models (HMM) to this task (Brugnara et al., 1993; Raphael, 1999; Vogel et al., 1996), leveraging the progress made in ASR systems with these models.

In the HMM framework, the events to align $\mathbf{y} = \{\mathbf{y}_0, \dots, \mathbf{y}_m, \dots, \mathbf{y}_{M-1}\}$ are manipulated via probabilistic *hidden states* $\mathbf{h} = \{\mathbf{h}_0, \dots, \mathbf{h}_t, \dots, \mathbf{h}_{T-1}\}$ such that $\forall t, \mathbf{h}_t \in \mathbf{y}$, that are not directly observable, but are expected to explain the audio frames $\mathbf{x} = \{\mathbf{x}_0, \dots, \mathbf{x}_t, \dots, \mathbf{x}_{T-1}\}$.

Determining the alignment between $\mathbf{x}$ and $\mathbf{y}$ through $\mathbf{h}$, requires to solve an inference problem, that is finding the most likely succession of events characterizing the audio. The formulation of the inference mechanism appears as a comprise between two important quantities: (1) a *prior evolution of hidden states* specifying, *e.g.*, their temporal evolution; and (2) an *observation model*, which measures a likelihood between audio observations $\mathbf{x}$ and the labels from the sequence $\mathbf{y}$ through the hidden states $\mathbf{h}$. The inference rule is expressed thanks to Bayes's theorem on conditional probabilities,

$$\underbrace{\mathbb{P}(\mathbf{h}|\mathbf{x})}_{\text{inference computation}} \propto \underbrace{\mathbb{P}(\mathbf{x}|\mathbf{h})}_{\text{observation model}} \underbrace{\mathbb{P}(\mathbf{h}).}_{\text{prior evolution}} \tag{3.19}$$

Note that since the quantity $\mathbb{P}(\mathbf{x})$, which should be part of the inference formula, is numerically independent of the hidden states to retrieve, it is usually dropped out of the computations.

As their name suggests, HMM assume the Markovian hypothesis, which stipulates that the audio at frame t (*i.e.*, $\mathbf{x}_t$) depends exclusively on its associated state (*i.e.*, $\mathbf{h}_t$) and, more drastically, that the sequence of states $\mathbf{h}$ is memoryless as well, such that

$$\mathbb{P}(\mathbf{x}|\mathbf{h}) = \mathbb{P}(\mathbf{x}_0, \dots, \mathbf{x}_{T-1}|\mathbf{h}_0, \dots, \mathbf{h}_{T-1}) = \prod_{t=0}^{T-1} \mathbb{P}(\mathbf{x}_t|\mathbf{h}_t). \tag{3.20}$$

This assumption is highly disputable but is widely accepted in practice as it allows the inference mechanism to remain tractable all along.

The inferred values precisely correspond to the timestamped encoding of the audio such that, according to Figure 3.2, $\mathcal{M}(\mathbf{x}) = \mathbb{P}(\mathbf{h}|\mathbf{x})$, and a decoding algorithm $\mathcal{D}_\pi$ (like Viterbi's) can be applied on it to force-align the true labels $\mathbf{y}$ to the audio $\mathbf{x}$.

Extensive literature can be found on methods for designing coherent prior evolution of hidden states $\mathbb{P}(\mathbf{h})$, *e.g.*, in score following (Cuvillier, 2016; Cuvillier and Cont, 2014), as well as defining robust observation models $\mathbb{P}(\mathbf{x}|\mathbf{h})$. The later emission probabilities are typically based on Gaussian Mixture Models (GMM) to compute a distance between prototypical and true audio features (Cont, 2009; Ljolje and Riley, 1991; Placeway and Lafferty, 1996; Raphael, 2006).

Such HMM-GMM aligners are ubiquitous in the voice alignment literature – *e.g.*, (Duan et al., 2013; Gong et al., 2015; Gorman et al., 2011; Hosom, 2009; McAuliffe et al., 2017; Rosenfelder et al., 2017).

Section summary – Traditional approaches to voice alignment

Dynamic Time Warping (DTW) – or VITERBI's decoding – is a general sequence-to-sequence alignment algorithm which constructs, recursively, an accumulative cost matrix based on local distances and permitted transitions between the sequences (*forward pass*) and retraces their alignment path from it (*backward pass*). In the light of its success, DTW (or a close variant) is a common choice for the decoding module $\mathcal{D}_\pi$. The acoustic modeling $\mathcal{M}$, on the contrary, has known various implementations, notably through Hidden Markov Models (HMM) coupled with Gaussian Mixture Models (GMM) that, by their probabilistic nature, could handle uncertainties when modeling the sequential structure of voice signals, despite assuming the disputable MARKOVIAN hypothesis.

3.4 Neural approaches to voice alignment

With the advent and success of deep learning in many fields – their data-driven approach outperforming various states of the art –, an increasing number of audio and voice-related tasks started adopting Deep Neural Networks (DNN) in their design. This is notably the case of the Automatic Speech Recognition (ASR) community in the 2010s (MOHAMED, 2014).

At first, DNN were integrated in hybrid systems in which they were coupled with Hidden Markov Models (HMM) . In these dual DNN-HMM architecture, DNN had the role to overcome the limitations of Gaussian Mixture Models (GMM) that were struggling to model complex data with high dimensions (MOHAMED et al., 2011). Then, end-to-end architectures were introduced so that HMM were completely removed from the ASR systems. In doing so, voice acoustic modeling was freed from the MARKOV assumption (GRAVES et al., 2013) and could benefit from the DNN capacity to capture long-term relationships in the voice signals. In opposition to earlier HMM-based models, end-to-end DNN-based systems are completely free from specialized and often complex expert knowledge (HANNUN et al., 2014). As a result, their data processing pipeline, training and inference usage are much simpler to handle than traditional or specialized approaches, yet not consistently better: the state of the art in lyrics transcription/alignment does integrate expert knowledge (GAO et al., 2021; GUPTA et al., 2020).

Given the proximity between voice recognition and voice alignment problems, the success of DNN reported in the ASR literature motivated their progressive adoption for voice-to-text synchronization. In such frameworks, the acoustic modeling function $\mathcal{M} \equiv \mathcal{M}_\Theta$ is parametrized by neural weights Θ, and the resulting timestamped encoding, or posteriorgram, $\mathcal{P} = \mathcal{M}_\Theta(\mathbf{x})$ can be fed to the decoding module $\mathcal{D}_\pi$, always referring to Figure 3.2.

As explained in section 3.1, the design of a training procedure implies the choice of loss function $\mathscr{L}(\Theta)$. Several approaches, based on different losses, for achieving forced alignment with DNN have been proposed in complement to standard methods and are thus presented.

3.4.1 Frame-wise classification

Early attempts in the deep learning-based voice alignment direction relied on an intuitive strategy consisting in letting the model predict one label per time step and ensuring that the correct symbol is effectively emitted at each instant. From this point of view, the alignment challenge is a frame-wise classification problem.

In this regard, by predicting *and* supervising symbol emission at the frame level, *both* the model outputs $\hat{\mathbf{y}} = \mathcal{M}_\Theta(\mathbf{x})$ *and* the ground truth reference $\mathbf{y}$ must be of the same length as the audio (*i.e.*, $M = T$), by means of duplicating all elements of the target sequence such that the duplicated elements align with the audio. The loss function, as generally presented in Eq. (3.7), to minimize for such a multi-category classification task, is based on a distance $\mathbb{d}$ known as the Categorical Cross-Entropy (CCE) such that

$$\mathscr{L}(\Theta) = \sum_{(\mathbf{x},\mathbf{y})\in\mathscr{D}} \mathbb{d}\left[\hat{\mathbf{y}}||\mathbf{y}\right] = \sum_{(\mathbf{x},\mathbf{y})\in\mathscr{D}} -\mathbf{y}\log(\hat{\mathbf{y}}) = \sum_{(\mathbf{x},\mathbf{y})\in\mathscr{D}} \left[-\sum_{t=0}^{T-1}\mathbf{y}_t\log(\hat{\mathbf{y}}_t)\right]. \tag{3.21}$$

Although great performances are to be denoted with this strategy (Backstrom et al., 2019; Kelley and Tucker, 2018), such training procedure calls for data in the form of paired audio and *already* precisely time-aligned text. This is a major shortcoming since only few accessible datasets provide that level of annotations, thus the data diversity and contexts seen by the model is intrinsically limited − as the process of manual labelling is tedious and extremely time-consuming.

Therefore, different approaches that could be trained with more widespread data, such as paired audio and text without precise timing information − that are available for, *e.g.*, Text-to-Speech (TTS) algorithms −, were investigated.

3.4.2 Attention mechanism

Schulze-Forster et al. (2020) precisely developed a system permitting the alignment between audios and their respective phonetic transcripts without requiring timestamps.

With the objective to isolate clean speech even in very noisy conditions, the authors proposed an encoder-decoder architecture based on Bidirectional LSTM (Bi-LSTM) and integrating a key attention mechanism (see section 3.1.3).

By taking both the audio $\mathbf{x}$ of length T and non-aligned sequences $\mathbf{y}$ of length M as inputs, the learnable weights involved in the attention mechanism $\boldsymbol{\alpha} \in [0,1]^{T\times M}$ turned out to resemble the accumulative score matrix intervening in the forward pass of the DTW algorithm (see section 3.3.1). An alignment path could thus directly be derived from $\boldsymbol{\alpha}$ through a backward pass.

The similar procedure have been applied to singing voice, however, due to its more challenging nature than speech (cf. section 2.2.3), the attention weights $\boldsymbol{\alpha}$ needed to be constrained to follow a monotonic path prior to any backward pass.

The authors proposed to directly impose a DTW-alike progression in the attention weights, thus introducing a new "DTW attention mechanism" (Schulze-Forster et al., 2021).

Summing up, voice-to-text alignment has been obtained as a positive side effect of voice separation, minimizing during training a simple loss function $\mathscr{L}(\Theta)$ measuring that the extracted vocals were close to the true vocals by means of a classical L1 distance choice for d.

3.4.3 Connectionist Temporal Classification (CTC)

Another approach for achieving voice alignment without the need of hard-labelled audio-symbol pairs was inspired from advances in end-to-end Automatic Speech Recognition (ASR) systems with a specialized cost function known as Connectionist Temporal Classification (CTC) .

The CTC loss was initially introduced to train Recurrent Neural Networks (RNN) on unsegmented data (Graves et al., 2006) and became, ever since, tremendously popular in the speech recognition literature (Collobert et al., 2016; Graves and Jaitly, 2014; Hori et al., 2017; Kim et al., 2017; Watanabe et al., 2017; Zhang et al., 2017).

Given some audio inputs $\mathbf{x}$, a CTC-trained neural network $\mathcal{M}_\Theta$ generates a *posteriorgram* $\mathcal{P} = \mathcal{M}_\Theta(\mathbf{x})$ which takes the form of per-frame discrete probability distributions over a finite alphabet of labels (*e.g.*, graphemes, phonemes, notes). In a speech recognition context, the temporal information contained in the posteriorgram is usually discarded as only an estimate $\hat{\mathbf{y}}$ of the target sequence $\mathbf{y}$ must be predicted but, yet again, this timestamping can explicitly be leveraged to force-align a ground-truth sequence $\mathbf{y}$ to the audio via a DTW-alike decoding algorithm $\mathcal{D}_\pi$. Concretely, CTC-based architectures have paved the way for the development of new, end-to-end alignment systems.

Regarding the alignment of *singing* utterances, and to the best of the author's knowledge, Stoller et al. (2019) were the first to explore such a CTC strategy and exploit the temporal axis of posteriorgrams to align raw audio waveforms, with background music, to their text transcripts. The acoustic modeling of the audio samples was a Wav-U-Net (Stoller et al., 2018) trained on a private dataset of more than 40k songs. Pursuing along the same trend, Vaglio et al. (2020a) proved that aligning singing voice with CTC could be extended to a multilingual context, even for languages with almost no training data, along with the robustness of CTC acoustic modeling for various tasks (Renault et al., 2021; Vaglio et al., 2020b, 2021).

Regarding the alignment of *speech* utterances, Kürzinger et al. (2020) showed that a CTC-based model, originally trained for ASR, could outperform legacy HMM-based methods for the automatic segmentation of large speech corpora into small audio excerpts and sentence-level texts.

In the light of its flexibility, the alignment systems proposed in this work are built upon a CTC strategy. Such a framework, though, has its own mathematical specificities and a decoding algorithm close but different to the classical DTW. The next section is dedicated to these notions.

Section summary – Neural approaches to voice alignment

Deep learning-based voice alignment consists in approximating the acoustic modeling function with a neural network $\mathcal{M}_\Theta$, parameterized by Θ, capable of extracting high-level and powerful features from the audio. Among the existing approaches found in the literature, *i.e.*, frame-wise classification (requiring hard labelling, *i.e.*, time-stamped ground truth sequences) and attention mechanisms (requiring an auxiliary task, *e.g.*, voice separation), the Connectionist Temporal Classification (CTC) is particularly appealing as it has shown great potential for end-to-end voice-to-sequence alignment with soft labelling. As a result, CTC is at the core of the acoustic models developed in this thesis.

3.5 CTC-based voice alignment

In the applicative context of end-to-end voice-to-sequence alignment, acoustic models trained to minimize the Connectionist Temporal Classification (CTC) loss are a pertinent choice. Proposed by Graves et al. (2006) in order to train a neural network for labelling unsegmented data, CTC introduced its specific formalism whose main definitions and original ideas relevant to this research are summarized in this section.

This background requires, on one hand, to detail the original CTC modeling $\mathcal{M}_\Theta$ for sequence transcription and, on the other hand, to present the CTC adaptation of a classic, DTW-based decoding module $\mathcal{D}_\pi$ for deriving alignment from the timing information of CTC predictions.

3.5.1 CTC-based acoustic modeling

Let $\mathcal{A}$ be a finite alphabet of symbols of size L associated with the set of sequences $\mathcal{A}^*$ such that

$$\mathbf{y} = \{\mathbf{y}_m\}_{m\in\{0,\dots,M-1\}} \in \mathcal{A}^* \iff \forall m \in \{0,\dots,M-1\}, \mathbf{y}_m \in \mathcal{A}.$$

Elements of $\mathcal{A}^*$ are referred to as *labellings*. Let $\mathcal{A}_\varepsilon$ denote the alphabet $\mathcal{A}$ extended with a blank label ε, and $\mathcal{A}_\varepsilon^*$ be the set of all sequences made of labels from $\mathcal{A}$ and blanks ε. Elements of $\mathcal{A}_\varepsilon^*$ are referred to as *(labelling) extensions*.

A neural network $\mathcal{M}_\Theta$ supervized by CTC typically receives as input a sequence $\mathbf{x} \in \mathbb{R}^{T\times F}$ from which it generates a posteriorgram $\mathcal{P} \in [0,1]^{T\times(L+1)}$. The posteriorgram results from the application of a softmax activation, according to Eq. (3.11), on the last dimension, such that each of its frames $\mathcal{P}[t], t \in \{0,\dots,T-1\}$ represents the emission probability distribution over the alphabet $\mathcal{A}_\varepsilon$, *i.e.*, over $L+1$ symbols including the blank label ε.

Consequently, $\mathcal{P}[t,l] \in [0,1]$ can be interpreted as the probability that the label $l \in \mathcal{A}_\varepsilon$ is emitted at the time instant t in the input sequence $\mathbf{x}$.

The probability of observing a labelling extension $\mathbf{l} \in \mathcal{A}_\varepsilon^*$ of length T, $\mathbf{l} = \{\mathbf{l}_0, \dots, \mathbf{l}_{T-1}\}$, conditionally on the inputs $\mathbf{x}$ associated with posteriorgram $\mathcal{P}$, is expressed as

$$\mathbb{P}(\mathbf{l}|\mathbf{x}) \quad = \quad \prod_{t=0}^{T-1} \mathcal{P}[t, \mathbf{l}_t]. \tag{3.22}$$

Let $\mathcal{B} : \mathcal{A}_\varepsilon^* \to \mathcal{A}^*$ denote the CTC mapping operator which, for any extension $\mathbf{l} \in \mathcal{A}_\varepsilon^*$, first merges all successively repeated symbols into one and then discards the blank labels ε, *e.g.*,

$\mathbf{l} \in \mathcal{A}_\varepsilon^*$ for $T = 15$	merge repetitions	discard blanks	$\mathcal{B}(\mathbf{l}) \in \mathcal{A}^*$
hεeeeeεεllεlεoo	hεeεlεlεo	hello	hello
hhhhεεεelεεlloo	hεelεlo	hello	hello
εεhheeεllεlεoεε	εheεlεlεoε	hello	hello

The previous examples illustrate the importance of the blank label ε is the CTC framework. Not only does it allow to separate the labels, even handling successive identical ones (*e.g.*, "ll" in "hello"), but also does not appear in the final predictions such that ε offers flexibility while constructing the posteriorgram $\mathcal{P}$. Due to its non-informative nature, embodying the choice not to specify any symbol from $\mathcal{A}$ at a given frame, predicting high probabilities for the blank label is an easy option for the network during its learning phase to explore without hard penalization (because blanks are not in the final output in opposition to a wrongly recognized symbol).

The operator $\mathcal{B}$ thus performs a many-to-one mapping as the same labelling $\mathbf{y} \in \mathcal{A}^*$ of size M can be obtained from several extensions $\mathbf{l} \in \mathcal{B}^{-1}(\mathbf{y})$ of size $T \geq M$. The conditional probability of observing the labelling $\mathbf{y}$ given the audio $\mathbf{x}$ is therefore the sum of *all* its associated extensions, *i.e.*,

$$\mathbb{P}(\mathbf{y}|\mathbf{x}) \quad = \quad \sum_{\mathbf{l} \in \mathcal{B}^{-1}(\mathbf{y})} \mathbb{P}(\mathbf{l}|\mathbf{x}). \tag{3.23}$$

A CTC-trained neural network aims to predict the labelling $\hat{\mathbf{y}} \in \mathcal{A}^*$ maximizing Eq. (3.23),

$$\hat{\mathbf{y}} \quad = \quad \underset{\mathbf{y} \in \mathcal{A}^*}{\operatorname{argmax}}\, \mathbb{P}(\mathbf{y}|\mathbf{x}) \tag{3.24}$$

or, which is equivalent, mimizing the associated negative-log-likelihood, which defines the CTC training criterion:

$$\hat{\mathbf{y}} \quad = \quad \underset{\mathbf{y} \in \mathcal{A}^*}{\operatorname{argmin}}\, -\log\mathbb{P}(\mathbf{y}|\mathbf{x}). \tag{3.25}$$

Such an optimization algorithm, very much alike DTW (see section 3.3.1), is intractable if approached naively – the cardinal of $\mathcal{B}^{-1}(\mathbf{y})$ being huge for a given labelling $\mathbf{y}$. Here again, Dynamic Programming (DP) techniques must be used to compute the CTC loss function, which is

$$\mathscr{L}_{\text{CTC}}(\Theta) \quad = \quad \sum_{(\mathbf{x},\mathbf{y}) \in \mathscr{D}} -\log\mathbb{P}(\mathbf{y}|\mathbf{x};\Theta) \quad = \quad \sum_{(\mathbf{x},\mathbf{y}) \in \mathscr{D}} -\log\left[\sum_{\mathbf{l} \in \mathcal{B}^{-1}(\mathbf{y})} \mathbb{P}(\mathbf{l}|\mathbf{x};\Theta)\right]. \tag{3.26}$$

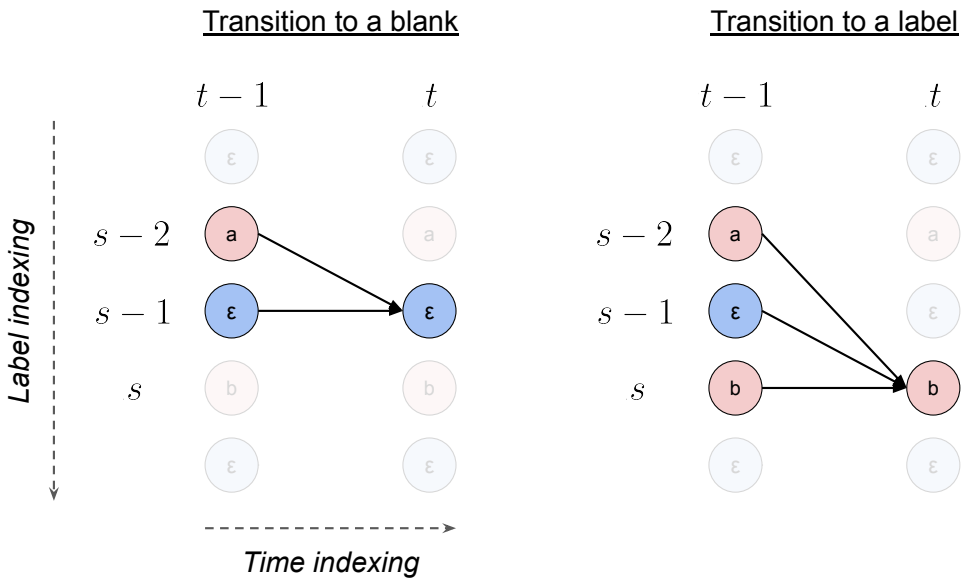


Figure 3.3: Transitions allowed in the CTC framework between labels and blanks, for the toy example target sequence "ab" with interleaved blanks ε.

In the specific CTC framework, however, the recursion rule for efficiently computing the probabilities is a bit different than DTW due to the introduction and major role of the blank label ε. Indeed, the possible value of the tth element $\mathbf{l}_t$ of an extension $\mathbf{l} \in \mathcal{A}_\varepsilon^*$ (associated with a labelling $\mathbf{y} \in \mathcal{A}^*$) depends on the previous element $\mathbf{l}_{t-1}$:

- When coming from a blank ε, an extension can either stay on the same blank ε or transition to a label $\mathbf{y}_m$, *i.e.*,

$$\mathbf{l}_{t-1} = \varepsilon \implies \mathbf{l}_t \in \{\varepsilon, \mathbf{y}_m\}. \tag{3.27}$$

- When coming from a label $\mathbf{y}_{m-1}$, an extension can either stay on the same label $\mathbf{y}_{m-1}$, transition to a blank label ε or jump directly to the next label $\mathbf{y}_m$, *i.e.*,

$$\mathbf{l}_{t-1} = \mathbf{y}_{m-1} \implies \mathbf{l}_t \in \{\mathbf{y}_{m-1}, \varepsilon, \mathbf{y}_m\}. \tag{3.28}$$

In order to take into account these transitions between blanks and non-blank labels, illustrated in Figure 3.3, Graves et al. (2006) defined a new sequence from $\mathbf{y}$, by adding a blank at the beginning and the end of it and interleaving blanks between every label. This sequence, denoted $\tilde{\mathbf{y}} = \{\tilde{\mathbf{y}}_s\}_{s\in\{0,\dots,S-1\}}$, has therefore a length $S = 2M + 1$, and

$$\tilde{\mathbf{y}} = \{\varepsilon, \mathbf{y}_0, \varepsilon, \dots, \varepsilon, \mathbf{y}_m, \varepsilon, \dots, \varepsilon, \mathbf{y}_{M-1}, \varepsilon\}. \tag{3.29}$$

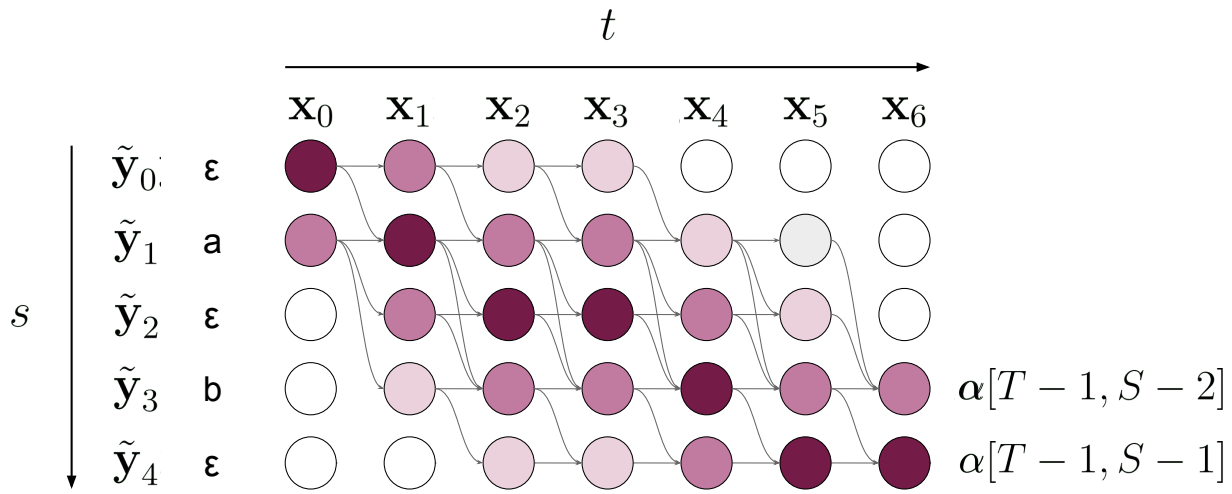


Figure 3.4: CTC loss computation via Dynamic Programming (DP) for the toy example "ab". Adaptation of the figure proposed by Hannun (2017).

This extended sequence is used to compute $\mathbb{P}(\mathbf{l}|\mathbf{x})$ dynamically with a reasoning resembling DTW – *i.e.*, through an accumulative score matrix $\boldsymbol{\alpha}$ between the (prefixes of the) sequences $\tilde{\mathbf{y}}$ and $\mathbf{x}$. Given the permitted transitions from Eq. (3.27) and Eq. (3.28), and as illustrated in Figure 3.4, $\boldsymbol{\alpha} \in \mathbb{R}^{T \times S}$ is computed following

$$\begin{aligned} &\forall t \in \{1, \dots, T-1\}, \forall s \in \{2, \dots, S-1\} \\ &\boldsymbol{\alpha}[t,s] = \mathcal{P}\left[t, \tilde{\mathbf{y}}_s\right] \times \begin{cases} \boldsymbol{\alpha}[t-1,s] + \boldsymbol{\alpha}[t-1,s-1] & \text{if } \tilde{\mathbf{y}}_s \in \{\varepsilon, \tilde{\mathbf{y}}_{s-2}\} \\ \boldsymbol{\alpha}[t-1,s] + \boldsymbol{\alpha}[t-1,s-1] + \boldsymbol{\alpha}[t-1,s-2] & \text{otherwise} \end{cases} \end{aligned} \tag{3.30}$$

with the initialization

$$\begin{aligned} \boldsymbol{\alpha}[t,0] &= 0 \quad \forall t \geq 0 \\ \boldsymbol{\alpha}[0,s] &= 0 \quad \forall s \geq 0. \end{aligned} \tag{3.31}$$

Finally, considering that a given extension $\mathbf{l}$ can terminate either with the final blank (index $s = S-1$) or with the last label (index $s = S-2$), the probability of observing $\mathbf{l}$ conditionally on $\mathbf{x}$ becomes:

$$\mathbb{P}(\mathbf{l}|\mathbf{x}) = \boldsymbol{\alpha}(T-1, S-1) + \boldsymbol{\alpha}(T-1, S-2). \tag{3.32}$$

As $\mathbb{P}(\mathbf{l}|\mathbf{x})$ depends on the posteriorgram $\mathcal{P}$, which is learned, this quantity is thus differentiable with respect to the model's learnable weights Θ and back-propagation algorithms can be used for training (Graves et al., 2006) and minimizing the loss $\mathscr{L}_{\text{CTC}}(\Theta)$ for, notably, ASR purposes.

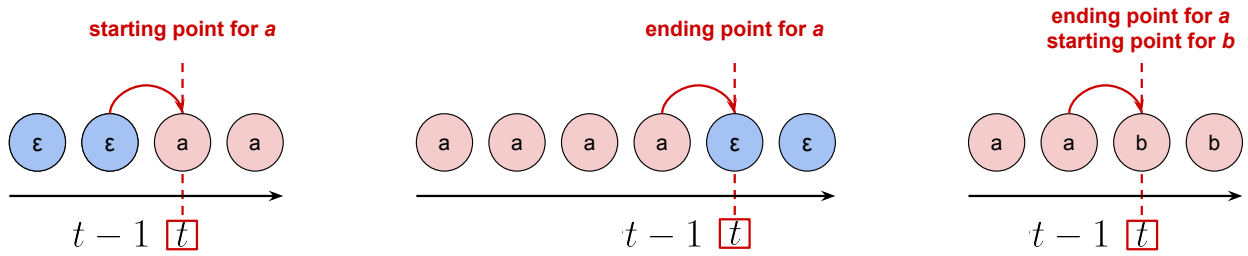


Figure 3.5: Alignment retrieval from the best decoded extension via the exploitation of transitions between two symbols or between symbol and blank.

3.5.2 CTC-based decoding module

The above-mentioned matrix $\boldsymbol{\alpha}$ is fundamental to compute the CTC loss and allow model training. Naturally, in this objective, $\boldsymbol{\alpha}$ produces a *marginalization over all possible alignment paths* − *i.e.*, all extensions $\mathbf{l} \in \mathcal{B}^{-1}(\mathbf{y})$ associated with the labelling $\mathbf{y}$ −, which is embodied by the sum operation of Eq. (3.30).

However, by considering the maximum operation instead of a summation, hence *maximizing* the probability $\mathbb{P}(\mathbf{l}|\mathbf{x})$, one can come back to a forced alignment scenario close to VITERBI's decoding. The initialization from Eq. (3.31) is unchanged, but the recursion rules become:

$$\begin{aligned} &\forall t \in \{1, \dots, T-1\}, \forall s \in \{2, \dots, S-1\} \\ &\boldsymbol{\alpha}[t,s] = \mathcal{P}\left[t, \tilde{\mathbf{y}}_s\right] \times \begin{cases} \max \begin{cases} \boldsymbol{\alpha}[t-1,s] \\ \boldsymbol{\alpha}[t-1,s-1] \end{cases} & \text{if } \tilde{\mathbf{y}}_s \in \{\varepsilon, \tilde{\mathbf{y}}_{s-2}\} \\ \max \begin{cases} \boldsymbol{\alpha}[t-1,s] \\ \boldsymbol{\alpha}[t-1,s-1] \\ \boldsymbol{\alpha}[t-1,s-2] \end{cases} & \text{otherwise.} \end{cases} \end{aligned} \tag{3.33}$$

This corresponds to the *forward pass*. Then, through the *backward pass*, one can retrieve the optimal extension $\mathbf{l}^*$, expressed as

$$\mathbf{l}^* = \underset{\mathbf{l} \in \mathcal{B}^{-1}(\mathbf{y})}{\operatorname{argmax}} \mathbb{P}(\mathbf{l}|\mathbf{x}) \tag{3.34}$$

by keeping trace of successive transitions retained during the forward pass.

Finally, the alignment path $\boldsymbol{\pi}^*$ between $\mathbf{x}$ and $\mathbf{y}$ is obtained by removing the blanks from $\mathbf{l}^*$ according to the visualizations in Figure 3.5, which gives insights on how start and end times for each symbol from $\mathbf{y}$ (*i.e.*, without blanks) are retrieved. This defines the CTC variant of the forced alignment decoding module $\mathcal{D}_\pi$ that will be used in this work.

Section summary – CTC-based neural alignment

Connectionist Temporal Classification (CTC) is a great option for end-to-end, soft-labelled neural voice alignment between an audio $\mathbf{x} \in \mathbb{R}^{T\times F}$ and a sequence $\mathbf{y} \in \mathcal{A}^*$ defined via an alphabet $\mathcal{A}$ of L symbols. The first step is a CTC-based acoustic modeling with a neural network $\mathcal{M}_\Theta$, which generates a posteriorgram $\mathcal{P} = \mathcal{M}_\Theta(\mathbf{x}) \in [0,1]^{T\times(L+1)}$ representing per-frame emission probabilities over the alphabet $\mathcal{A}$ and a blank label ε. This extra token, ε, is responsible for the flexibility of a CTC approach, as the model can choose not to specify any symbol from $\mathcal{A}$ but must eventually recognize the target sequence $\mathbf{y}$ from $\mathcal{P}$. The second step is the decoding module $\mathcal{D}_\pi$ exploiting the temporal information in $\mathcal{P}$ towards final alignment retrieval. This module is implemented as a variant of the classical VITERBI's algorithm in a CTC context, *i.e.*, accounting for the additional transitions and subtleties induced by the introduction of the blank ε.

3.6 Scientific background for voice alignment in a nutshell

Chapter summary – Scientific background for voice alignment

In this chapter, I introduced the definitions and mathematical tools necessary to tackle the voice alignment problem between an audio $\mathbf{x} \in \mathbb{R}^{T\times F}$ and a sequence $\mathbf{y} \in \mathcal{A}^*$ whose symbols belong to a finite alphabet $\mathcal{A}$. I showed that any synchronization system is based on an acoustic model $\mathcal{M}$ and a decoding module $\mathcal{D}_\pi$. In this thesis, exploiting recent advances in deep learning for voice processing and their robustness through their ability to learn from data themselves, the acoustic models are implemented as Deep Neural Networks (DNN) parameterized by Θ, *i.e.*, $\mathcal{M} \equiv \mathcal{M}_\Theta$. I motivated the choice of a Connectionist Temporal Classification (CTC) framework for the acoustic model $\mathcal{M}_\Theta$ due to its great practical flexibility, enabling model training with only soft labelling and the design of end-to-end architectures. Following such an approach, the alignment between an audio $\mathbf{x}$ and a sequence $\mathbf{y} \in \mathcal{A}^*$ relies on (1) a *posteriorgram* $\mathcal{P} = \mathcal{M}_\Theta(\mathbf{x})$ representing per-frame probabilities over the alphabet $\mathcal{A}$ and the CTC non-informative blank label ε; and (2) a variant of the DTW algorithm, accounting precisely for the blank ε, which allows the forced alignment *per se*.

Part II

Contributions: time-constrained neural voice alignment

Chapter 4

ADAGIO: An acoustic model for temporal voice alignment

*"Scientific research involves going beyond the well-trodden and well-tested ideas and theories that form the core of scientific knowledge. During the **Time** scientists are working things out, some results will be right, and others will be wrong. Over **Time**, the right results will emerge."*

— Lisa Randall

This chapter aims to present **ADAGIO** – the acoustic model at the heart of this thesis work which serves as a base for the defended contributions. Beyond the acronym standing for **A**utomatic **D**eep **A**li**G**nment of v**OI**ce, ADAGIO is an end-to-end model thought for voice-to-symbol alignment and developed within the CTC framework.

As that is the case for many advances in research, ADAGIO is based on several inspirations from the literature. Therefore, the section 4.1 exposes the neural architectures used for the acoustic modeling of recent alignment systems mentioned in the previous chapter and retraces the early explorations that rose the fundamental requirements that an acoustic model should, in the current context, satisfy. In light of these considerations and thought process, the proper design and (fixed) parameters of ADAGIO are detailed in section 4.2. A summary of this chapter is given in section 4.3.

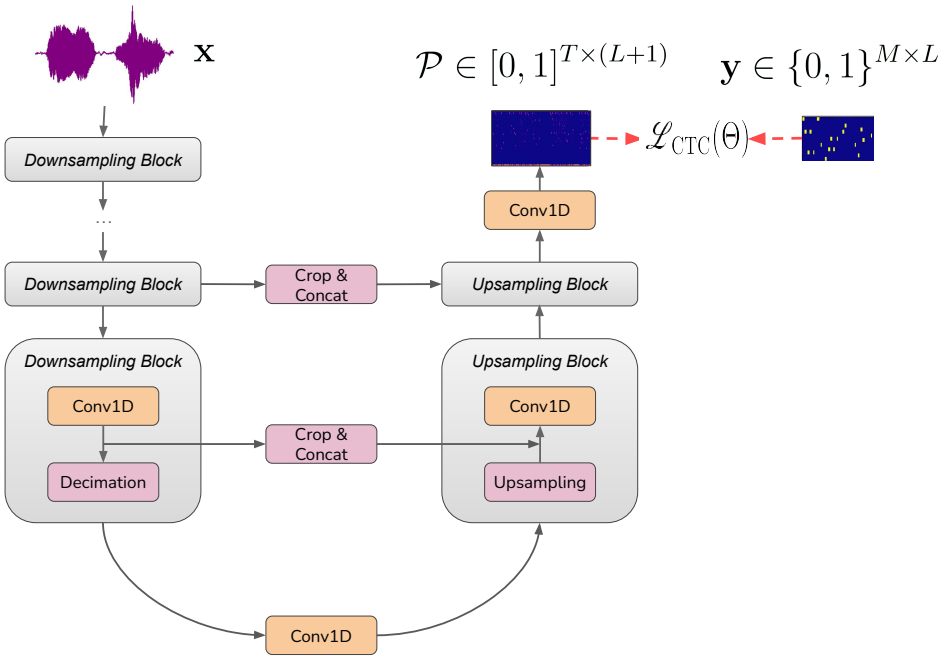


Figure 4.1: The Wav-U-Net architecture for CTC-based voice alignment (Stoller et al., 2019).

4.1 Model history

Before introducing the ADAGIO system itself, this section elaborates the stages of its design, notably by tracing the inspirations from the models in the literature and the first experimental tests carried out. In this perspective, the neural baselines – as anchors and comparison points – are presented and discussed. The first developmental drafts, run in the context of phonetic alignment with clean signals, are specified. ADAGIO being born from their extension, and especially from the overcoming of their intrinsic limits, these are exposed.

4.1.1 Baselines

First and foremost, the existing systems for addressing voice alignment are presented as they constitute inspirations (clarified in the next section 4.1.2) and baselines for this work.

Wav-U-Net

Stoller et al. (2019) implemented the acoustic model as a CTC-trained Wav-U-Net whose architecture is shown in Figure 4.1.

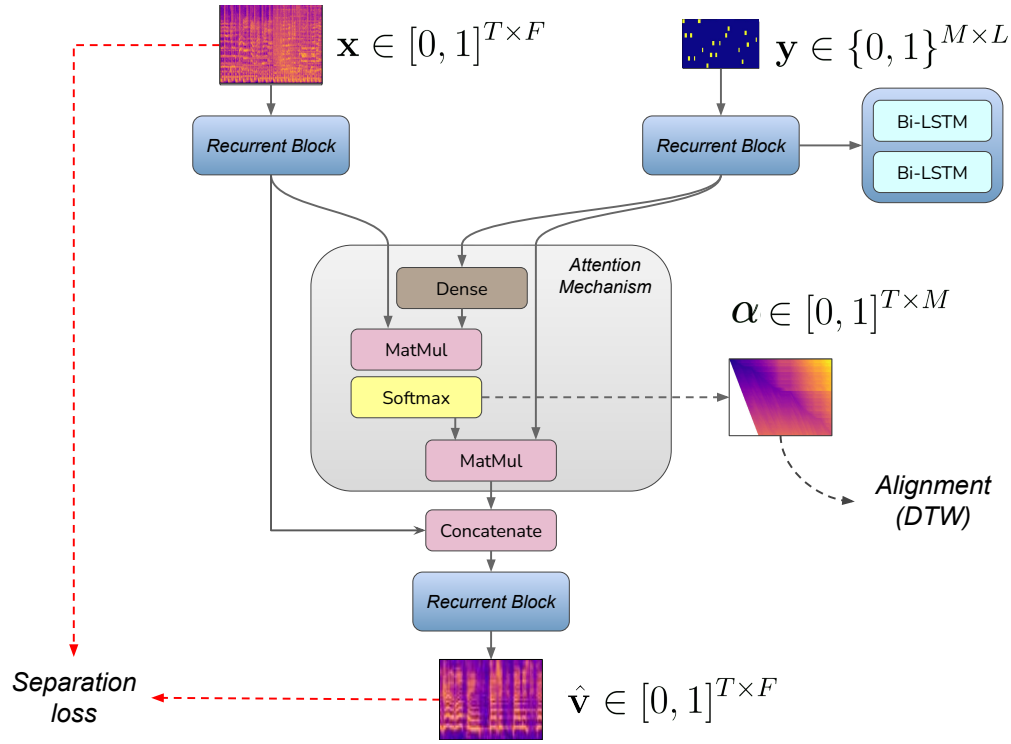


Figure 4.2: The recurrent-attention (ARNN) architecture for separation-based voice alignment (Schulze-Forster et al., 2021).

It processes raw audio recordings, *i.e.*, 1-D signals, through a series of 1D convolution and downsampling layers allowing to capture low-level and higher level features from the audio at multiple time resolution. To generate the posteriorgram, to be optimized with respect to the CTC criterion, upsampling and more 1D convolutions are used. The core of any U-Net architecture is to propagate a multi-level modeling via concatenations between the downsampling and upsampling blocks.

Attentional Recurrent Neural Network (ARNN)

Schulze-Forster et al. (2021) implemented the acoustic model as an Attentional Recurrent Neural Network (ARNN) whose architecture is shown in Figure 4.2.

It processes *both* text and audio with recurrent layers and an attention mechanism. The model supervises the reconstruction of the (denoised/clean) voice inputs from the output of the attention layer. The alignment can be derived from the same attention mechanism, as it exploits *attention weights* measuring the "weighting" of each symbol in each audio frame, which is in line with a typical Dynamic Time Warping (DTW) accumulative score matrix.

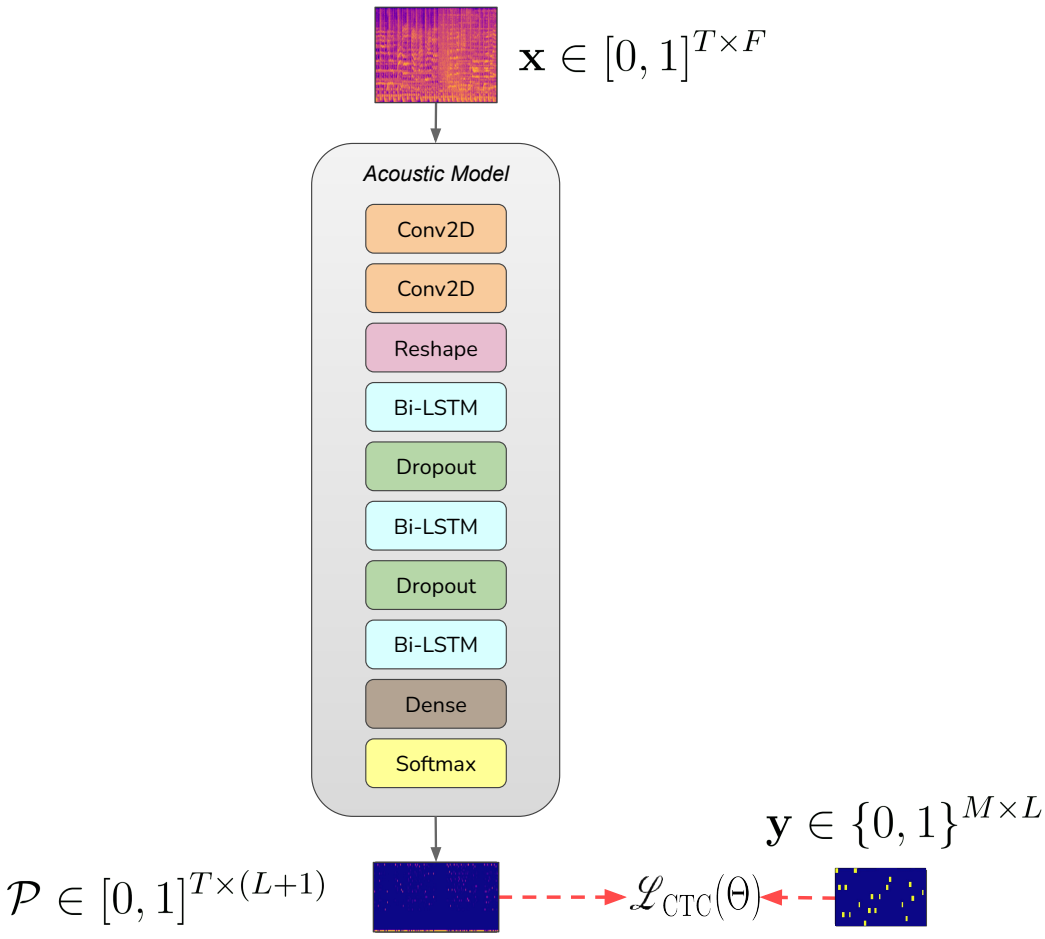


Figure 4.3: The convolutional-recurrent (CRNN) architecture for CTC-based voice alignment (Vaglio et al., 2020a).

Convolutional Recurrent Neural Network (CRNN)

Vaglio et al. (2020a) implemented the acoustic model as a plain Convolutional Recurrent Neural Network (CRNN) whose architecture is shown in Figure 4.3.

It processes log-mel-spectrograms first with two convolutional layers, extracting more abstract features from the audio representation, then with three recurrent layers (Bi-LSTM) with dropout in-between, and a final (time-distributed) dense layer which generates the desired posteriorgram. The training procedure is supervised by the CTC loss.

4.1.2 Early explorations and learnings

First attempt

The first proposal made in this thesis was highly inspired from the architecture of Schulze-Forster et al. (2021) where the separation loss was replaced by the CTC training criterion. Hence, the defined acoustic model was exploiting both non-aligned transcripts $\mathbf{y}$ and audios $\mathbf{x}$ as inputs of the acoustic model and was benefiting from an attention mechanism to generate relevant posteriorgram $\mathcal{P}$. This model was well suited for phonetic alignment, *i.e.*, alignment of an audio recording with its associated sequence of pronounced phonemes. In the case of clean speech and solo singing voice, the feasibility to develop an end-to-end CTC-based voice aligner was confirmed, even with high temporal precision (as required by the phonetic granularity). This architecture was the starting point of the publication (Teytaut and Roebel, 2021).

Second attempt

However, there was room for improvements:

- On one hand, it was desired to get free from the symbols $\mathbf{y}$ as second input of the model for two reasons:
  - First, because the attention mechanism involves a softmax that generates, for each of the T frames, a probability distribution over the M symbols, *i.e.*, $\boldsymbol{\alpha} \in [0,1]^{T\times M}$ in Figure 4.2. During training, only short audio-text pairs are fed to the network due to memory constraints − hence, the attention weights $\boldsymbol{\alpha}$ are specialized on small sequences. But in concrete inference usage, *e.g.*, aligning a song with its lyrics, the softmax operation would be over a much greater number M of symbols than during the training phase, resulting in blurred and unexploitable attention weights $\boldsymbol{\alpha}$. A solution would be to pre-segment the whole recording into shorter audio-symbol pairs, inducing circular dependency, which is neither straightforward nor convenient. This approach was, therefore, not suitable for a real world inference scenario.
  - Second, because having the symbols as inputs is fundamentally preventing the model from being used for transcription. Although it is not the purpose of this work, predicting and evaluating transcriptions from audio remains an interesting option.
- On the other hand, Recurrent Neural Networks (RNN) are computationally intensive and may be tricky to train. As an associated trend in the ASR literature, it was aimed to replace such layers. The increasingly popular Transformer architecture (Vaswani et al., 2017), and its application to speech recognition with CTC (Miao et al., 2020), highlighted that similar performances could be obtained without recurrent layers (and less parameters) thanks to simpler dense layers coupled with self-attention mechanisms.

Following these observations, the Bi-LSTM were replaced by Conv 1D layers, which capture a wider temporal context than pure dense layers (Collobert et al., 2016). The attention mechanism became a multi-head self-attention (Vaswani et al., 2017) and, as now independent from the target sequence, the overall model could be used for aligning whole recordings − on condition of applying it successively on small audio portions and concatenate the outputs. This second design was the starting point of the publication (Teytaut et al., 2022).

Limitations

This second approach, though, also remained limited and needed to be improved. Although applicable on entire audio recordings, the attention mechanism was still imposing a segmentation stage (of the audio only). For convenience, the acoustic model should be usable in one-shot on voice signals. Second, and more importantly, this model was small and focused on the phonetic alignment tasks with only short and clean speech and singing voice databases − without noises or background music −, which were, once again, not representative of real world inference scenario.

4.1.3 Acoustic model requirements

Given these explorations, insights were gained on the several requirement that any robust and practical acoustic model should satisfy. There are summarized in this section.

At *training* time,

— Parsimony is desirable in terms of number of parameters for the network and data neediness for the training procedure *per se*;

— Non-Recurrent architectures are preferred to Recurrent Neural Networks (RNN) and their variants (and even more so for CTC-based alignment, as detailed later).

At *inference* time,

— Audio-Only models, in terms of inputs, are straightforward to apply and therefore recommanded;

— Polyphonic usage is essential as the less pre-processing of the audio, the better. The system must be robust to musical accompaniment and usable on real recordings directly, without, *e.g.*, relying on a source separator to isolate vocals.

None of the existing literature proposals satisfy all of these criteria. Indeed, the data-intensive nature of the Wav-U-Net has required its training on a private dataset of more than 40k songs, much more than any publicly available dataset, and its large architecture that cannot fit on limited hardware infrastructure. The ARNN and CRNN both exploit recurrent layers.

		Wav-U-Net	CRNN	ARNN	ADAGIO
Training	Parsimony		✓	✓	✓
	Non-Recurrent	✓			✓
Inference	Audio-Only	✓	✓		✓
	Polyphonic	✓		✓	✓

Table 4.1: Advisable criteria for a robust voice-to-symbol alignment system.

The CRNN uses Spleeter (Hennequin et al., 2020) as a pre-processing step to extract vocals for both training and inference, while the ARNN needs the target sequence as second input.

The Table 4.1 summarizes these observations. The core proposal of this research is to develop a new system in line with all of the identified criteria – ADAGIO.

Section summary – Model history

The early explorations of the current research, which has been based on and inspired from the existing models of the literature – Wav-U-Net (Stoller et al., 2019), CRNN (Vaglio et al., 2020a) and ARNN (Schulze-Forster et al., 2021) –, have allowed to identify criteria that any robust voice-to-symbol alignment system should satisfy but were not systemically fulfilled. These are: parsimony both in terms of data neediness and architecture size, absence of recurrent layers, audio-based inference only and robustness to noise/music accompaniment for real world application on, *e.g.*, polyphonic music. This thesis precisely aims to develop a system in line with these objectives – ADAGIO.

4.2 ADAGIO: Automatic Deep AliGnment of vOIce

This section is dedicated to the presentation of ADAGIO, the acoustic model at the center of this research. Based on convolutional architecture and the Connectionist Temporal Classification (CTC) framework, ADAGIO allows generating posteriorgrams characterizing the temporal evolution of the symbolic information contained in voice signals. The core architecture of the acoustic model is depicted in Figure 4.4.

Data pre-processing

The network takes as inputs normalized log-mel-spectrograms $\mathbf{x} \in [0,1]^{T \times F \times 1}$ that are derived from the audio. For numerical stability and uniformity over the whole dataset, the Time-Frequency Representations (TFR) are scaled between 0 and 1 based on a 80dB threshold, as commonly done in the community – *e.g.*, this is the default audio reading setting of the `librosa` audio library for Python (McFee et al., 2015).

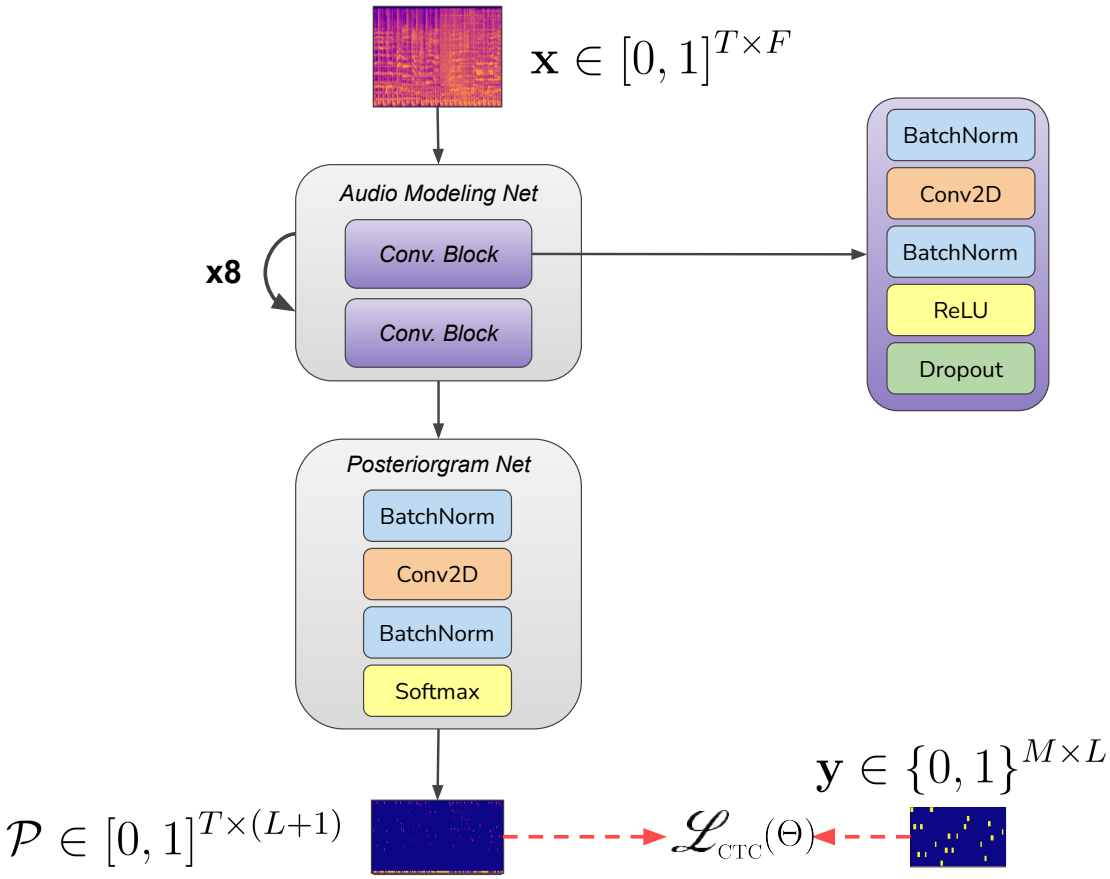


Figure 4.4: ADAGIO – Neural architecture.

The ground-truth transcripts $\mathbf{y}$ serve for supervision.

During the on-the-fly creation of the data batches, the TFR and symbolic sequences are padded if necessary so that all audios (resp. sequences) have the same number of frames T (resp. number of symbols M), *i.e.*, the greatest one. Zero-padding is used on raw audios and a dedicated padding token is used for the labels.

Model design

ADAGIO's neural layers are organized into two successive subnets:

— an *Audio Modeling Net*, in charge of capturing deep learned featured from the audio;

— and a *Posteriorgram Net*, which simply generates the posteriorgram from the learned feature map.

Net	Stage	Layer	Input size	Output size
Audio Modeling	$n = 1$	Conv. Block 1	$T \times F \times 1$	$T \times F \times C$
		Conv. Block 2	$T \times F \times C$	$T \times F/2 \times C$
	$n > 1$	Conv. Block 1	$T \times (F/2^{n-1}) \times (2^{n-2} \cdot C)$	$T \times (F/2^{n-1}) \times (2^{n-1} \cdot C)$
		Conv. Block 2	$T \times (F/2^{n-1}) \times (2^{n-1} \cdot C)$	$T \times (F/2^{n}) \times (2^{n-1} \cdot C)$
	$\vdots$	$\vdots$		
	$n = 8$	Conv. Block 1	$T \times 1 \times (2^{n-2} \cdot C)$	$T \times 1 \times E$
		Conv. Block 2	$T \times 1 \times E$	$T \times 1 \times E$
Posteriorgram		Batch Norm	$T \times 1 \times E$	$T \times 1 \times E$
		Conv 2D	$T \times 1 \times E$	$T \times 1 \times (L+1)$
		Batch Norm	$T \times 1 \times (L+1)$	$T \times 1 \times (L+1)$
		Softmax	$T \times 1 \times (L+1)$	$T \times 1 \times (L+1)$

Table 4.2: ADAGIO – summary of neural layers and input-output shapes.

The *Audio Modeling Net* is a fully convolutional network composed of 8 successive stages of 2 convolution blocks. A convolutional block is constructed with the following layers: batch normalization, 2D-convolution, batch normalization, ReLU activation, and dropout. No time pooling is performed – the number of frames T is thus completely determined by the signal processing setup. At each stage, the first block uses a 1×1 stride but the second one uses a 1×2 stride for the 2D convolution, hence halving the number of Mel frequency bins, and the number of convolutional filters (common to both blocks) is increased by two times. By changing the number of frequency bins and channels every other layer, more convolutional blocks can be integrated, increasing the modeling power and the *receptive field*.

The *Posteriorgram Net* is composed of a final succession of batch normalization, a last 2D convolution, batch normalization and a softmax activation to obtain the per-frame probability distributions over the labels. The number of channels of the last convolution is the desirable dimension for the posteriorgram, *i.e.*, the size L of the symbol alphabet $\mathcal{A}$ plus the blank token. Predictions can be derived from the posteriorgram $\mathcal{P} \in [0, 1]^{T \times (L+1)}$.

For each layer, the shapes of the input-output tensors are shown in Table 4.2 and the network hyperparameters are specificed in Table 4.3. Note that, with this setup, the global receptive field of the 16 convolutional layers is 1024ms. Therefore, the model can be trained on audio excerpts of several seconds (typically 10–20s), but can be used in inference on audio with much longer durations thanks to its fully convolutional architecture.

Parameter	Notation	Value	Unit	Remark
Signal processing				
Sampling rate	F_S	16000	Hz	
Window nature	$w[n]$			HANN window – see Eq. (2.10)
Window length		1024	samples	
FFT size		1024		
Hop size	H	512	samples	
Frame duration	δt	32	ms	$= 1000 \times H/F_S$
Top dB value		80	dB	Audio normalization threshold
Mel bins	F	128		
Neural architecture				
Filters (init.)	C	16		
Filters (max.)	E	512		
Kernel size	K	3×3		
Deep depth	D	8		Number of stages of the *Audio Modeling Net*
Receptive field		1024	ms	$= 2 \times D \times (K-1) \times \delta t$

Table 4.3: ADAGIO – hyperparameter setup.

Section summary – ADAGIO: Automatic Deep AliGnment of vOIce

ADAGIO is an end-to-end, fully convolutional network which processes normalized audio log-mel-spectrograms $\mathbf{x} \in [0,1]^{T \times F \times 1}$. Its neural architecture consists of 16 convolutional blocks proceeding to an audio modeling with a receptive field around 1s, and a final convolutional block generating the posteriorgram $\mathcal{P} \in [0,1]^{T \times (L+1)}$ over the L labels and the CTC blank ε.

4.3 Acoustic model for temporal voice alignment in a nutshell

Chapter summary – An acoustic model for temporal voice alignment

In this chapter, I presented ADAGIO, the acoustic model at the core of my thesis, which is dedicated to the **A**utomatic **D**eep **A**li**G**nment of v**OI**ce. Retracing the early explorations of my research, I shed light on relevance and convenience criteria that a robust voice aligner should satisfy. For the sake of flexibility, and according to these criteria, ADAGIO has been implemented as an end-to-end, fully convolutional network trainable with a reasonable (*i.e.*, publicly available) amount of data and directly applicable on voice signals in the presence of music accompaniment.

Chapter 5

Temporal constraints for alignment enhancement

*"One must work with **Time** and not against it."*

– Ursula Kroeber Le Guin

It has been seen that a temporal voice alignment system depends on an acoustic model and a decoding module. The previous chapter, exploiting the technical tools of deep learning, was dedicated to the introduction of ADAGIO – a convolutional neural network predicting time-symbol posteriorgrams from audio that can be exploited to force-align a target sequence. The quality of the temporal information contained in the posteriorgram is thus essential and crucial for the relevance of the estimated alignments. The purpose of this chapter precisely is to reinforce the temporality of the neural predictions with ADAGIO via temporal constraints.

The section 5.1 first exposes why ADAGIO, due to the Connectionist Temporal Classification (CTC) formalism itself, carries some *limitations for alignment*, which motivates a search for a better robustness. The fundamental section 5.2 introduces proposals to this aim. Concretely, temporal (1) spectral reconstruction, (2) structure propagation and (3) guided monotony are thought as *additive temporal constraints*. These ideas are practically implemented as loss functions to be minimized on the vocals in addition to CTC, so that a *multi-objective training* phase is specified in section 5.3. In the end, this new *time-constrained* version of ADAGIO is fully summarized in section 5.4 prior to a chapter recap in section 5.5.

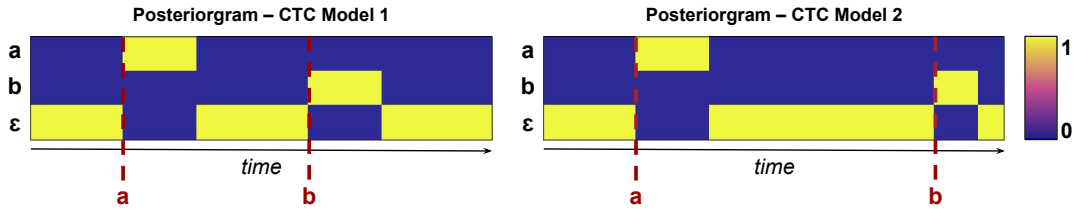


Figure 5.1: The CTC alignment limitation: Two posteriorgrams both relevant for *transcription* of the text "ab" but leading to fundamentally different *alignments*.

5.1 The need for additional constraints

Previous Chapter 4 introduced ADAGIO as an acoustic model built upon the Connectionist Temporal Classification (CTC) algorithm. In section 3.5, the formalism of CTC was thoroughly presented. Its founding principle is an one-to-many prediction framework based on the existence of a blank label ε that allows, by its non-informative nature, to decode several acceptable sequences from a given input (Graves et al., 2006; Hannun, 2017). This is achieved by inserting blanks ε between the symbols of the target sequence $\mathbf{y}$ – creating the so-called extensions $\mathbf{l}$ of $\mathbf{y}$. By *marginalizing over* all possible extensions $\mathbf{l}$ from $\mathbf{y}$ and interleaved blanks, a CTC-trained neural network eventually estimates a conditional probability of $\mathbf{y}$ given $\mathbf{x}$ – *i.e.*, a posteriorgram $\mathcal{P}$ that can be further used to force-align a sequence of symbols.

However, this approach also means that CTC cannot guarantee accurate temporal alignments between the input and output sequences, by the very nature of this one-to-many mapping it exploits. The blank symbols can occur at any time step in the output sequence, which makes it difficult to know precisely when each audio frame is aligned with each symbol.

The Figure 5.1 illustrates this on a toy example where two different CTC-trained models produce different yet relevant posteriorgrams as both of them (1) perfectly *recognize* the target text "ab"; but (2) lead to drastically different temporal alignments. Such time shifts were often empirically observed during this thesis and were naturally dependent on the neural architecture at play (*e.g.*, recurrences or convolutional receptive field). This can be problematic for some applications, such as automatic captioning of videos, where the timing of the output text is crucial and, in a wider scope, for temporal audio and voice alignments in general.

This is the reason why speech specialists started introducing attention mechanisms as Eq. (3.14) in joint CTC-attention models. Attention allows handling relevant dependencies between audio and symbols over time, which can benefit downstream tasks as speech recognition (Kim et al., 2017; Park et al., 2022; Watanabe et al., 2017). Extending such approaches for designing voice synchronizers precisely resembles Schulze-Forster et al. (2021)'s proposal, at the cost of (1) increasing computation complexity; and (2) being dependent on the target sequence for the acoustic modeling.

Some approaches have tried regularization on the CTC loss to better understand the role and utmost importance of its blank label (Bluche et al., 2015), prevent peaky probability distribution (Liu et al., 2018), or improve its scalability with Cross-Entropy (CE) through sampling (Variani et al., 2018). But, all in all, and to the best of the author's knowledge, no existing work has been dedicated to explicitly address and ensure the emergence of temporal alignment directly in CTC probabilities.

This thesis has focused on this point since its very beginning, alongside the search of the neural architecture for ADAGIO, as seen in the two publications (Teytaut and Roebel, 2021; Teytaut et al., 2022) on which this chapter is based on. The final proposals inheriting from these explorations are detailed in the next sections to temporally constraint CTC posteriorgrams.

Section summary – The need for additional constraints

Connectionist Temporal Classification (CTC) has launched a new trend in the voice alignment literature with the great benefit of not requiring aligned data for training deep acoustic models. Yet, alignment remains intrinsically difficult to couple with CTC. Indeed, by nature, CTC measure a *transcription* cost and can therefore be minimized without guaranteeing precise *alignment* properties. There thus exists a need to define additional constraints to reinforce the temporality in the CTC predictions.

5.2 Temporal constraints for reinforcing alignment

This section details the temporal constraints proposed to enhance the emergence of voice alignment properties in CTC posteriorgrams, which has been one of the key contributions of this thesis work – in addition to the basic neural architecture of ADAGIO, which serves as a starting point hereinafter.

The expression "temporal constraints" convey two essential pieces of information. First, the term "*contraints*" is used to mean that the ramifications added to the initial network take the form of additional *cost functions* that will also be minimized during neural training alongside the CTC. Second, the term "*temporal*" clearly states that these introduced losses are concerned with the temporal axis of involved tensors, in order to precisely bring in temporal knowledge during the optimization phase.

In this context, three angles to define new temporal training objective have been identified. The first one is to ensure an accurate temporal reconstruction of the spectral content. The second goal is to preserve the structural organization from the audio recordings. The third angle aims to guarantee the occurrence of a precise monotonicity between the predictions of the acoustic model and the symbolic sequence to be decoded through a classical accumulative score. This section explores each of these research paths. All details about the related network structures can be found in Table 5.1 at the end of this chapter.

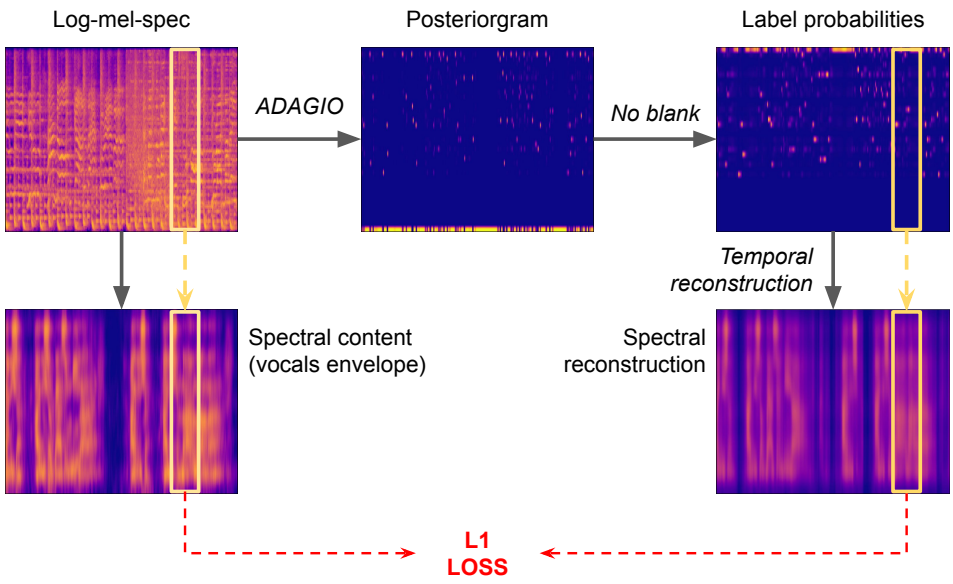


Figure 5.2: Illustration of the proposed spectral reconstruction content.

5.2.1 Spectral reconstruction

The first approach proposed to help the model predicting the symbols at their accurate and relevant position is a supplementary constraint that consists in reconstructing the audio spectral information from the CTC posteriorgram, which can be seen as a compressed representation of the audio. The reconstruction is performed by a small neural network − the *Reconstruction Net*. It is also of convolutional nature − this way, the spectral reconstruction at a given frame t depends only on a small temporal context around t in the posteriorgram. The inputs to the *Reconstruction Net* are the probabilities of all symbols without the blank as it is precisely desired to limit the importance of the blank in the overall prediction as it is mostly responsible for the CTC intrinsic alignment limitation (see section 5.1). During training, the supervision is done with a L1 distance loss between the estimated and real spectral content. The Figure 5.2 illustrates this additional reconstruction constraint in the case of voice-to-text alignment.

Preliminary research

In the first study (Teytaut and Roebel, 2021), the reconstruction of the entire input spectrogram $\mathbf{x} \in [0,1]^{T \times F}$ was supervised. In the second study (Teytaut et al., 2022), it was stated that such a systematic global reconstruction might not be pertinent. For instance, the propagation of the F0 and harmonics is irrelevant for the alignment of phonemes − or any textual modality of communication more generally − as phonemes are associated with formants.

Thus, the network would have to dedicate some of its modeling capacity to produce an F0 estimation that would not be pertinent for the alignment task. In this second study, the spectral reconstruction was supervised on the Mel-Frequency Cepstral Coefficients (MFCCs). In this manuscript, it is proposed to directly rely on the *spectral envelope* that fully characterizes formants and is a F0-free Time-Frequency Representations (TFR) – see section 2.2.2.

Vocals separation

However, considering the spectral envelope of the entire audio mixture is no interest as the envelope associated with musical accompaniment conveys no information on the phonetic content. Therefore, the spectral reconstruction must only be based on the *vocals* contained in the audio recordings. This implies the usage of a source separation algorithm, a common need in the voice community (Vincent et al., 2018), and here specifically a singing voice extractor (Jansson et al., 2017).

A re-implementation of Choi et al. (2019)'s neural voice separator was at disposal upon training with data augmentation (Cohen-Hadria et al., 2019; Lancaster and Souviraà-Labastie, 2020). It achieves high quality extraction of the voice and these estimated vocals were judged clean enough to be reliable for constraining the network. More information can be found on a dedicated section 7.3.2 where the voice separation algorithm is detailed, evaluated (Vincent et al., 2006) and integrated in a musicological pipeline.

Note that the isolated vocals solely serve during the training phase. They are *not* required for inference: only the audio mixture shall be fed to the network.

Final proposal

Given an estimation $\hat{\mathbf{e}} \in [0,1]^{T\times F}$ of the "true" *spectral envelope* $\mathbf{e} \in [0,1]^{T\times F}$ extracted from solo vocals upon voice separation, the spectral reconstruction loss $\mathscr{L}_{\text{REC}}(\Theta)$ is defined as the following L1 distance:

$$\mathscr{L}_{\text{REC}}(\Theta) \;=\; \|\hat{\mathbf{e}} - \mathbf{e}\|_1 \;=\; \sum_{t=0}^{T-1}\sum_{f=0}^{F-1} |\hat{\mathbf{e}}[t,f] - \mathbf{e}[t,f]| \,. \tag{5.1}$$

Figure 5.3 shows the ramification of ADAGIO associated with this spectral reconstruction constraint.

It is worth mentioning that this proposal remains valid for the case of *note* alignment although one should precisely consider spectral *excitation* – *i.e.*, Fundamental Frequency (F0) and harmonics or everything but the spectral envelope. Remembering the source-filter modeling in Eq. (2.15) and Figure 2.7, the information of the varying pitches will be well represented in the spectral excitation, which is was should be supervised in that context.

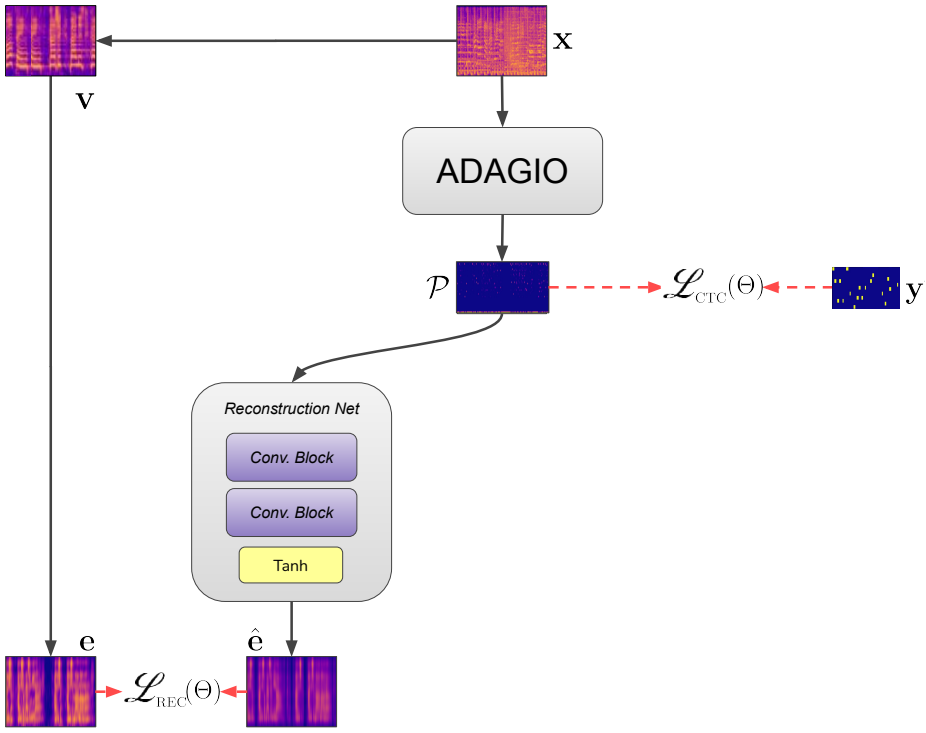


Figure 5.3: Spectral reconstruction constraint (here, for voice-to-*text* alignment).

5.2.2 Temporal structure propagation

The second approach proposes to study the *temporal structure* with the claim that shared similarity patterns, informing on the local temporal structure, are to be found in the original voice signals and in the alignment predictions. To do this, Self-Distance Matrix (SDM) are computed on the input audio – as features representing the structural content – and are to be recovered from the CTC posteriorgrams. This is performed by a small neural network – the *Structural Net*. It is also of convolutional in nature – this way, the estimated structure at a given frame t depends only on a small temporal context around t in the posteriorgram. The inputs to the *Structural Net* are again the probabilities of all symbols *without* the blank. During training, the supervision is done with a L1 distance loss between the estimated and real SDM. The Figure 5.4 illustrates this additional reconstruction constraint in the case of voice-to-text alignment.

Self-Distance Matrix (SDM)

Self-Distance Matrix (SDM) are well-known representations for capturing the structure of pieces of music (Cohen-Hadria and Peeters, 2017; Fell et al., 2022).

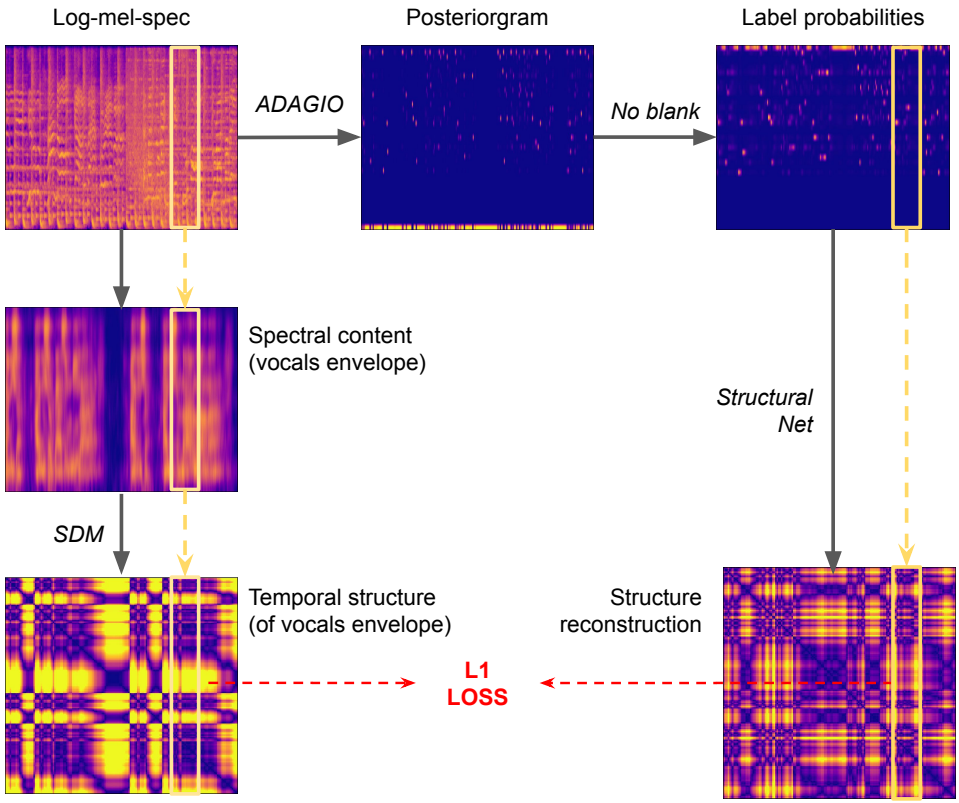


Figure 5.4: Illustration of the proposed temporal structure propagation.

The SDM $\mathcal{S} \in [0,1]^{T\times T}$ of the log-mel-spectrogram $\mathbf{x} \in [0,1]^{T\times F}$ measures a proxmitiy between every spectral frame. The value $\mathcal{S}[t,t'] \in [0,1]$ is the dissimilarity between the frame t and t'. The so-called dissimilarity implies the choice of a distance, whose nature depends on the studied context, calculated between the frame contents.

The lack of contrast in SDM is known for often limiting their discriminative power (Peeters and Angulo, 2022), which was commonly observed when relying on a cosine distance measure. For data whose value range is [0,1], a L1 distance based on their L2 norms coupled with a simple normalization factor has empirically appeared as a sufficiently discriminative choice.

Therefore, in this work, the non-resemblance between frames is defined as a frequency-based scoring based on L2 norms, *i.e.*,

$$\forall t,t' \quad \mathcal{S}[t,t'] \;=\; \frac{1}{F^p}\sum_{f=0}^{F-1}\left|\mathbf{x}[t,f]^2 - \mathbf{x}[t',f]^2\right|. \tag{5.2}$$

where $p = 1/\sqrt{2}$ has been fixed upon empirical observations.

For the same reasons as exposed before for spectral reconstruction, the structure of the musical background does not carry relevant information for voice alignment. Hence, the ground-truth structural information is derived from the spectral *envelope* (for text synchronization) or *excitation* (for notes synchronization) of the *solo vocals* estimated thanks to a voice separator.

Finally, as (1) one is looking for *local* patterns but not up to the *frame* level; and (2) all SSM are $(T \times T)$-shaped hence memory demanding, a (4×4)-average pooling operation with stride (2×2) is used both to smooth local structural singularities and reduce memory storage.

Comparisons with previous work

The structural loss was introduced in the publication (Teytaut et al., 2022), although Self-*Similarity* Matrices were considered in that paper (with no loss of purpose). They were computed using the cosine similarity on the complete log-mel-spectrogram, hence including both spectral envelope and source (F0 and harmonics), which was not an optimal strategy. The data already were clean solo speech and singing signals so the extension to real world recordings was limited.

Here, the reference structure is derived from the spectral envelope of the solo vocals estimated, once again, via a voice separator. This separation is only required for training and not inference. The other main difference between previous proposal and this manuscript regarding temporal structure is that the SDMs were computed directly on the posteriorgrams $\mathcal{P}$ in (Teytaut et al., 2022) while they are derived from an extra encoding from $\mathcal{P}$ here. This choice was motivated by realizing that the structural constraint alone was not helping the alignment procedure, and *ad hoc* analyses revealed that posteriorgrams could be trickily shaped to have correct structures, yet without predicting the full duration of each label. In order to prevent this negative side effect from happening, while continuing to propagate the temporal structure information, it is here proposed to generate a new feature map from $\mathcal{P}$ and compute the SDM of this one.

Final proposal

Given an estimation $\hat{\mathcal{S}} \in [0,1]^{\frac{T}{2} \times \frac{T}{2}}$ of the *temporal structure* $\mathcal{S} \in [0,1]^{\frac{T}{2} \times \frac{T}{2}}$ extracted from *solo vocals envelope* upon voice separation and after the strided average pooling (hence $T/2$ frames), the structural propagation loss $\mathscr{L}_{\text{STR}}(\Theta)$ is defined as the following L1 distance:

$$\mathscr{L}_{\text{STR}}(\Theta) \;=\; \left\| \hat{\mathcal{S}} - \mathcal{S} \right\|_1 \;=\; \sum_{t=0}^{\frac{T}{2}-1} \sum_{t'=0}^{\frac{T}{2}-1} \left| \hat{\mathcal{S}}[t,t'] - \mathcal{S}[t,t'] \right|. \tag{5.3}$$

The Figure 5.5 shows the ramification of ADAGIO associated with this structural propagation constraint.

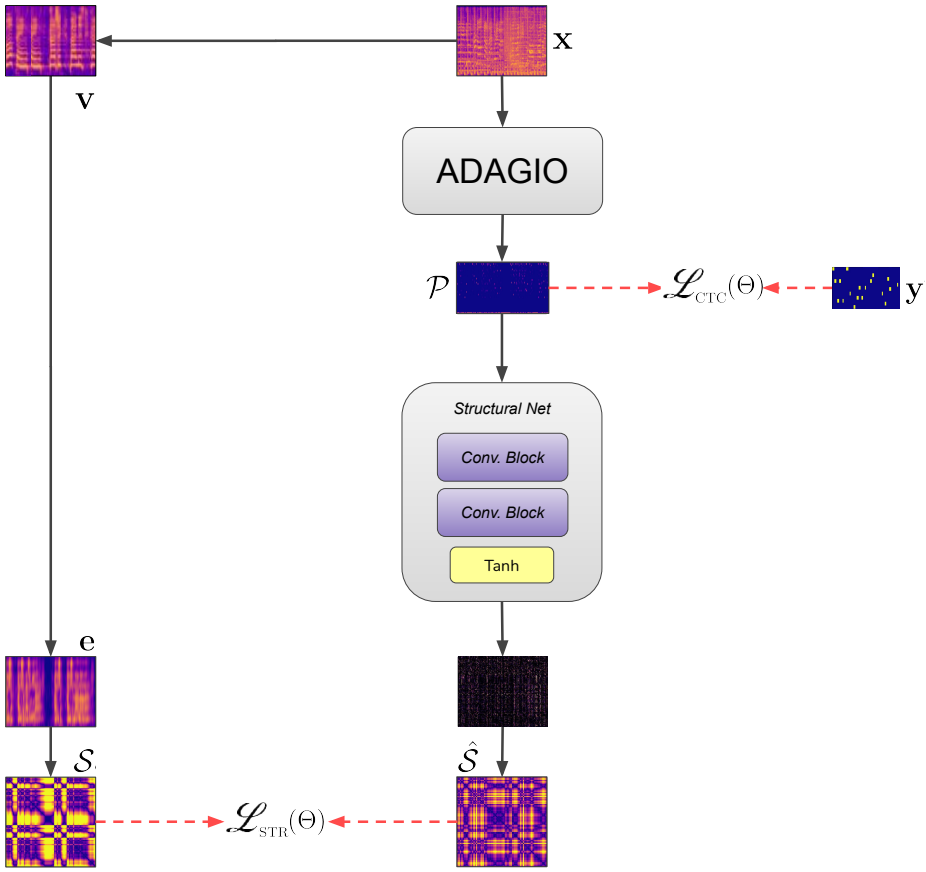


Figure 5.5: Structural propagation constraint (here, for voice-to-*text* alignment).

5.2.3 Guided audio-symbol monotony

The third and final approach proposes to directly measure and quantify a sequence-to-sequence accumulative score, typical derived from the Dynamic Time Warping (DTW) algorithm. Speech and singing signals and symbolic transcripts are ordered sequences. Remembering the very definition of alignment between general sequences (see section 2.3.1), synchronizing audio with such symbols implies uncovering a path with *monotonic* properties and relevant transitions between the symbols. This is performed by a small external module – the *Monotonic Net*. Although it is not a neural network *per se* (no extra parameters), it does define a new gradient to be propagated through the posteriorgram. The inputs to the *Monotonic Net* are the CTC symbol probabilities (without blank) and the target sequence. The symbolic sequence is thus used during training to supervise CTC and this new constraint but it is still not needed at inference time. A DTW-alike score is obtained and becomes the loss aimed at being minimized.

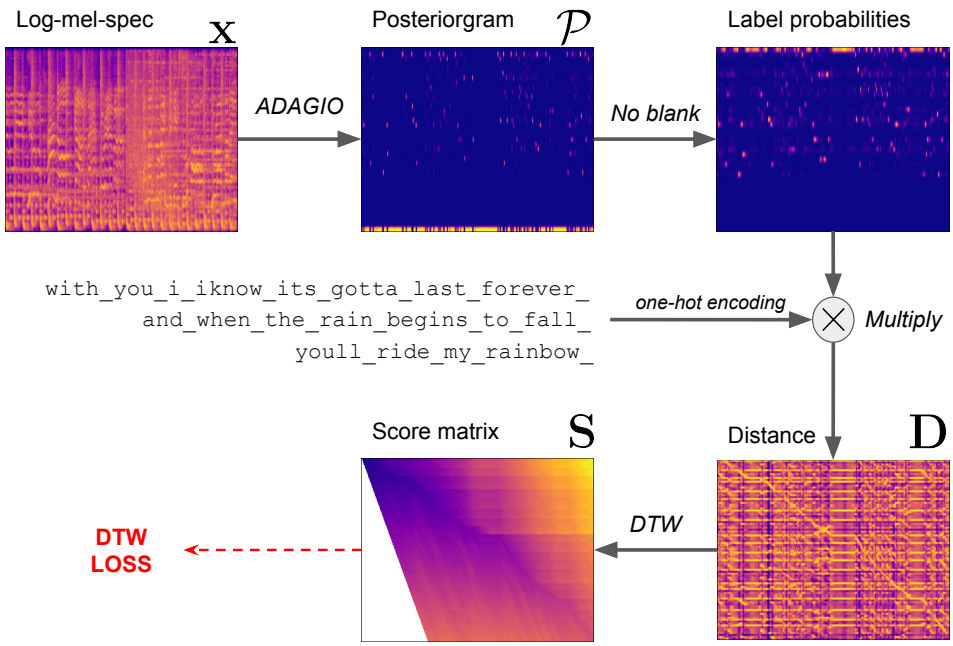


Figure 5.6: Illustration of the proposed time-symbol guided monotony.

The Figure 5.6 illustrates this additional audio-symbol monotonic constraint in the case of voice-to-text alignment.

Audio-symbol likelihood

The starting point of this additional constraint is the definition of a proximity measure between each of the T frames and each of the M symbols. Such a matrix is easy to obtain from the posteriorgram without blank $\mathcal{P} \in [0,1]^{T \times L}$ and the one-hot encoded symbols $\mathbf{y} \in \{0,1\}^{M \times L}$. From these two tensors, one can compute the audio-symbol log-likelihood $\mathbf{D} \in [0,1]^{T \times M}$ by taking the opposite of the logarithm of the direct matrix multiplication between $\mathcal{P}$ and $\mathbf{y}$. The probabilities are set to a minimum threshold of $1e^{-9}$ so that these log-likelihoods have an upper bound of 9Np (Neper), which allows renormalization of the measures between 0 and 1. In short, with T the transpose operator,

$$\mathbf{D} \quad = \quad -\log\left(\mathcal{P}\mathbf{y}^{\mathrm{T}}\right). \tag{5.4}$$

The existence of a monotonic alignment implies the emergence of a pseudo-*diagonal* in matrix $\mathbf{D}$ (hence its denomination), which can actually be seen in Figure 5.6. This diagonal is supposedly associated with the best alignment path to uncover.

Accumulative score matrix

The next step, highly similar to Dynamic Time Warping (DTW) from Eq. (3.17), consists in computing an accumulative score matrix with only two permitted transitions $\mathcal{G} = \{(1,1),(1,0)\}$ – *i.e.*, coming from the same label or the previous label at the previous frame. This accumulative score $\mathbf{S} \in \mathbb{R}^{T \times M}$ uses $\mathbf{D}$ as reference for distances. It follows:

$$\begin{array}{llrcl} t=0 & m=0 & \mathbf{S}[t,0] & = & \mathbf{D}[t,0] \\ t=0 & \forall m>0 & \mathbf{S}[0,m] & = & \infty \\ \forall t>0 & \forall m>0 & \mathbf{S}[t,m] & = & \mathbf{D}[t,m] \;+\; \min \begin{cases} \mathbf{S}[t-1,m] \\ \mathbf{S}[t-1,m-1]. \end{cases} \end{array} \tag{5.5}$$

Note that, as opposed to the DTW, which usually maximises a resemblance, the min operator is used here to define a training criterion to minimize.

Guided audio-symbol monotony

Following Eq. (3.18), the final accumulative value is a direct measure of the (mis)alignment between the two sequences, *i.e.*, between time and symbols, *e.g.*, between voice signals and text or notes. This leads to introduce the monotonic, DTW-alike loss $\mathscr{L}_{\text{DTW}}(\Theta)$ as:

$$\mathscr{L}_{\text{DTW}}(\Theta) \;=\; \mathbf{S}[T-1, M-1]. \tag{5.6}$$

By minimizing this score, the network is *guided* towards the best alignment path. This can happen only if CTC systematically highlights the full duration of each label and not only their onset. The Figure 5.7 shows the ramification of ADAGIO associated with this guided audio-symbol monotony constraint.

Comparison with previous work

The idea to exploit the sequence $\mathbf{y}$, which is available during training, to measure a time-label proximity based on the posteriorgram $\mathcal{P}$ has been introduced in (Teytaut et al., 2022). The matrix $\mathbf{D}$ was similarly defined and it was precisely desired to ensure the existence of a diagonal in it. For this, inspired from *guided* attention literature (Tachibana et al., 2018), a Gaussian-decreasing matrix was used to extract a fixed diagonal out of $\mathbf{D}$ and derive a loss value from it. However, this strategy carried a prior that all labels should have similar duration, which was intrinsically not true. The continuation of this work was rather to couple $\mathbf{D}$ with a better alignment quality metric – Dynamic Time Warping (DTW), hence the current proposal.

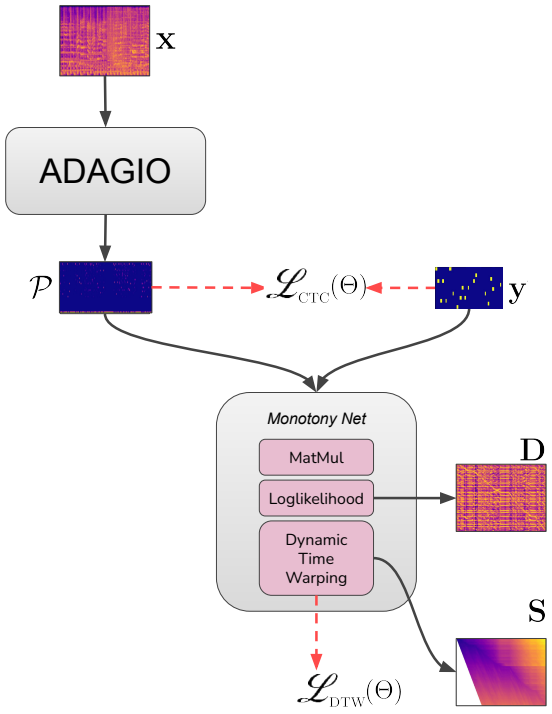


Figure 5.7: Guided monotony constraint (here, for voice-to-*text* alignment).

Section summary – Temporal constraints for reinforcing alignment

In order to reinforce both onset and duration detection of *non-blank* labels in CTC posteriorgrams $\mathcal{P}$, with the ultimate goal to guarantee the emergence of alignment properties directly in the probabilities, three temporal constraints were defined, namely:

- **Spectral reconstruction** aiming at reproducing the relevant spectral information (*e.g.*, *envelope* for text or *excitation* for note alignment) from the label probabilities;
- **Temporal structure propagation** ensuring that the same structural as the input spectral content can be temporally retrieved from the posteriorgrams;
- **Guided audio-symbol monotony** measuring an accumulative (mis)alignment score derived from both the labels and the posteriorgrams.

Vocals extraction is performed prior to the supervision of the two former cases as musical accompaniment is an inherent obstacle to learn the reconstruction and structure of *voice*.

5.3 Multi-objective training

The present research aims to couple various losses together − the CTC and three additional constraints − thus raising the question of how to combine them and to determine their respective trade-off, which is a concern for all multi-task learning problems (Kim et al., 2017; Liang et al., 2021b).

Since all the losses are duration-dependent, it is proposed to ensure that they all scale similarly, linearly with the time length T, which is the dimension of interest for temporal audio alignment. Prior to the scaling process itself, an idea of the relative order of magnitude of the different losses is necessary. They are estimated from worst-case scenario studies, and allow, in the end, to build a global loss function to minimize during training.

5.3.1 Worst-case scenario studies

This section presents the estimation of cost functions from critical and theoretical scenarios, which tend to represent the early stages of training where the weights are random and the network has not yet had time to specialize. For each loss, its associated formula is remembered and further investigated. It is reminded that $\mathbf{x} \in [0,1]^{T\times F}$ denotes the input log-mel-spectrogram and $\mathbf{y} \in \mathcal{A}^M$ or $\{0,1\}^{M\times L}$ represents the target symbols as a linear sequence or one-hot encoding. The model's learnable parameters are denoted Θ.

Connectionist Temporal Classification (CTC)

Retracing the definition of the Connectionist Temporal Classification (CTC) training criteria from Eq. (3.25), the CTC loss can be expressed as

$$\mathscr{L}_{\text{CTC}}(\Theta) \;=\; -\log\left(\sum_{\substack{\ell\in\mathbf{l}\\ \mathbf{l}\in\mathcal{B}^{-1}(\mathbf{y})}}\prod_{t=0}^{T-1}\mathcal{P}[t,\ell]\right). \tag{5.7}$$

The number of alignment paths considered in the sum of the CTC computation is the cardinal of $\mathcal{B}^{-1}(\mathbf{y})$, which contains all (labelling) extensions $\mathbf{l}$ that reduce to $\mathbf{y}$ after application of the operator $\mathcal{B}$ merging repeated character and removing blanks. Hannun (2017) and Mao (2019) revealed that there were $\binom{T+M}{T-M}$ possibilities. In the most uninformative way possible, the CTC posteriorgram would be uniform over the L labels and blank. An estimate of the loss value then becomes:

$$\mathscr{L}_{\text{CTC}}(\Theta) \;\sim\; -\log\left[\binom{T+M}{T-M}\left(\frac{1}{L+1}\right)^{T}\right]. \tag{5.8}$$

Going further, it can be assumed that there are much more time frames T than symbols to synchronize M, *i.e.*, $T \gg M$, leading to:

$$\mathscr{L}_{\text{CTC}}(\Theta) \quad \sim \quad \log(L+1)T. \tag{5.9}$$

Note that the CTC algorithm necessarily needs that $T \geq M$. (One cannot map each of the M symbols to an audio frame if there are less than M frames.)

Spectral reconstruction

As seen in Eq. (5.3), the spectral reconstruction directly compares an estimate of the spectral content, *e.g.*, envelope $\hat{\mathbf{e}} \in [0,1]^{T \times F}$ to a reference one $\mathbf{e} \in [0,1]^{T \times F}$ by minimizing:

$$\mathscr{L}_{\text{REC}}(\Theta) \quad = \quad \|\hat{\mathbf{e}} - \mathbf{e}\|_1 \quad = \quad \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} |\hat{\mathbf{e}}[t,f] - \mathbf{e}[t,f]| \,. \tag{5.10}$$

Given the value range and shapes associated with both tensors, the maximum difference one could theorically measure is:

$$\mathscr{L}_{\text{REC}}(\Theta) \quad \sim \quad FT. \tag{5.11}$$

Temporal structure propagation

Similarly, the temporal structure loss minimizes the difference between the input Self-Distance Matrix (SDM) $\mathcal{S} \in [0,1]^{\frac{T}{2} \times \frac{T}{2}}$ and an estimated one from the posteriorgram $\hat{\mathcal{S}} \in [0,1]^{\frac{T}{2} \times \frac{T}{2}}$. It comes:

$$\mathscr{L}_{\text{STR}}(\Theta) \quad = \quad \left\| \hat{\mathcal{S}} - \mathcal{S} \right\|_1 \quad = \quad \sum_{t=0}^{\frac{T}{2}-1} \sum_{t'=0}^{\frac{T}{2}-1} \left| \hat{\mathcal{S}}[t,t'] - \mathcal{S}[t,t'] \right| . \tag{5.12}$$

Given the value range and sizes associated with both tensors, the maximum difference one could theorically measure is:

$$\mathscr{L}_{\text{STR}}(\Theta) \quad \sim \quad \tfrac{1}{4}T^2. \tag{5.13}$$

Guided audio-symbol monotony

Finally, the monotony constraint is based on a per-frame accumulative penalty score, see Eq. (5.5). It is maximized when CTC does not recognize any label correctly, and even less their full duration. In this case, the greatest distance possible (which is 1) is added at each of the T frames. In the end:

$$\mathscr{L}_{\text{DTW}}(\Theta) \quad \sim \quad T. \tag{5.14}$$

5.3.2 Scaling the losses

It has been shown that chosen objective criteria do not result in similar variations when the audio length, *i.e.*, number of frames T, changes. This is a major issue since different elements in the training set will not end up inducing comparable updates for the gradients and weights themselves.

Consequently, from previous worst-case scenario studies, highlighting *estimates* of the loss value, it is proposed to scale each loss so that they all have identical, linear dependency on T, which is reasonable because time segments in sequences will always have the same impact independently of the phrase they are found in.

The scaled losses are defined as:

$$\begin{aligned}
\mathscr{L}^n_{\text{CTC}}(\Theta) &\leftarrow \frac{1}{\log(L+1)} \mathscr{L}_{\text{CTC}}(\Theta) \\
\mathscr{L}^n_{\text{REC}}(\Theta) &\leftarrow \frac{1}{F} \mathscr{L}_{\text{REC}}(\Theta) \\
\mathscr{L}^n_{\text{STR}}(\Theta) &\leftarrow \frac{4}{T} \mathscr{L}_{\text{STR}}(\Theta) \\
\mathscr{L}^n_{\text{DTW}}(\Theta) &\leftarrow \mathscr{L}_{\text{DTW}}(\Theta)
\end{aligned} \tag{5.15}$$

All in all, the global loss to be minimized during training is:

$$\mathscr{L}(\Theta) = \mathscr{L}^n_{\text{CTC}}(\Theta) + \tfrac{1}{3}\sum\nolimits_i s_i \mathscr{L}^n_i(\Theta) \tag{5.16}$$

with i an index over the above-mentioned constraints and $s_i \in \{0, 1\}$ whether the constraint i is used or not. The factor $\frac{1}{3}$ aims at preventing even all joint constraints from dominating the CTC, which remains the core objective as it does generate the posteriorgrams. This results in 8 possible configurations to be evaluated, which is precisely the core of the next Part III – to evaluate and apply ADAGIO as an acoustic modeling.

Section summary – Multi-objective training

The proposals to integrate more temporal information during the training of the acoustic model ADAGIO suffer from not having the same dependency on the number of audio frames, which is a problem for their direct combination. To cope with this, based on worst-case scenario studies, a scaling factor was estimated to normalize each of them, ensuring that the losses all have an indentical, linear variability with respect to audio length.

5.4 Time-constrained acoustic modeling

All in all, a *time-constrained* version of ADAGIO has been proposed and implemented to reinforce alignment quality. The Table 5.1 depicts the succession of layers and tensor shapes for ADAGIO and these extensions. It notably sheds light on the number of convolutional filters used for spectral reconstruction and structure propagation. These new convolutions use a kernel size of 1×1. The other hyperparameters are the same as the ones shared in Table 4.3.

Section summary – Time-constrained acoustic modeling

Summing up, three additional temporal constraints were proposed with the aim to reinforce ADAGIO by assuring the emergence of alignment properties directly in the posteriorgrams. Each of these supplementary constraints is implemented in practice by adding a subnetwork that is connected to the output of ADAGIO, *i.e.*, the CTC posteriorgram *without* blank. The respective *Reconstruction Net*, *Structural Net* and *Monotony Net* are therefore integrated to complement the original neural architecture. This *time-constrained* extension of ADAGIO is fully summarized in Figure 5.8.

5.5 Temporal constraints for alignment enhancement in a nutshell

Chapter summary – Temporal constraints for alignment enhancement

In this chapter, I presented a *time-constrained* extension of my acoustic model ADAGIO. I started by explaining why the Connectionist Temporal Classification (CTC) formalism, notably due to the blank label ε, cannot guarantee precise temporal alignment. I proposed to explicitly address this problem by introducing three additional temporal constraints aimed at forcing the posteriorgram to capture deep alignment properties by integrating time-dependent information. These contributions took the form of three auxiliary tasks of spectral reconstruction, temporal structure propagation, and guided time-sequence monotony. Each of them relied on a similar idea: to generate a new representation heavily dependent on coherently aligned emissions in the posteriorgram. They required completing the neural architecture and design a multi-objective training scenario by normalizing the diverse losses at play. The final upcoming chapters evaluate the relevance of these proposals and present their concrete applications to voice research.

Net	Stage	Layer	Input size	Output size
Audio Modeling	$n = 1$	Conv. Block 1	$T \times F \times 1$	$T \times F \times C$
		Conv. Block 2	$T \times F \times C$	$T \times F/2 \times C$
	$n > 1$	Conv. Block 1	$T \times (F/2^{n-1}) \times (2^{n-2} \cdot C)$	$T \times (F/2^{n-1}) \times (2^{n-1} \cdot C)$
		Conv. Block 2	$T \times (F/2^{n-1}) \times (2^{n-1} \cdot C)$	$T \times (F/2^{n}) \times (2^{n-1} \cdot C)$
	$\vdots$	$\vdots$		
	$n = 8$	Conv. Block 1	$T \times 1 \times (2^{n-2} \cdot C)$	$T \times 1 \times E$
		Conv. Block 2	$T \times 1 \times E$	$T \times 1 \times E$
Posteriorgram		Batch Norm	$T \times 1 \times E$	$T \times 1 \times E$
		Conv 2D	$T \times 1 \times E$	$T \times 1 \times (L+1)$
		Batch Norm	$T \times 1 \times (L+1)$	$T \times 1 \times (L+1)$
		Softmax	$T \times 1 \times (L+1)$	$T \times 1 \times (L+1)$
Reconstruction		Conv. Block 1	$T \times 1 \times L$	$T \times 1 \times E$
		Conv. Block 2	$T \times 1 \times E$	$T \times 1 \times F$
		Tanh	$T \times 1 \times F$	$T \times 1 \times F$
Structural		Conv. Block 1	$T \times 1 \times L$	$T \times 1 \times E$
		Conv. Block 2	$T \times 1 \times E$	$T \times 1 \times F$
		Tanh	$T \times 1 \times F$	$T \times 1 \times F$
Monotony		MatMul	❶ $T \times 1 \times L \equiv T \times L$ ❷ $M \times L$	$T \times M$
		Log-likelihood	$T \times M$	$T \times M$
		DTW	$T \times M$	$T \times M$

Table 5.1: Time-constrained ADAGIO − summary of neural layers and input-output shapes.

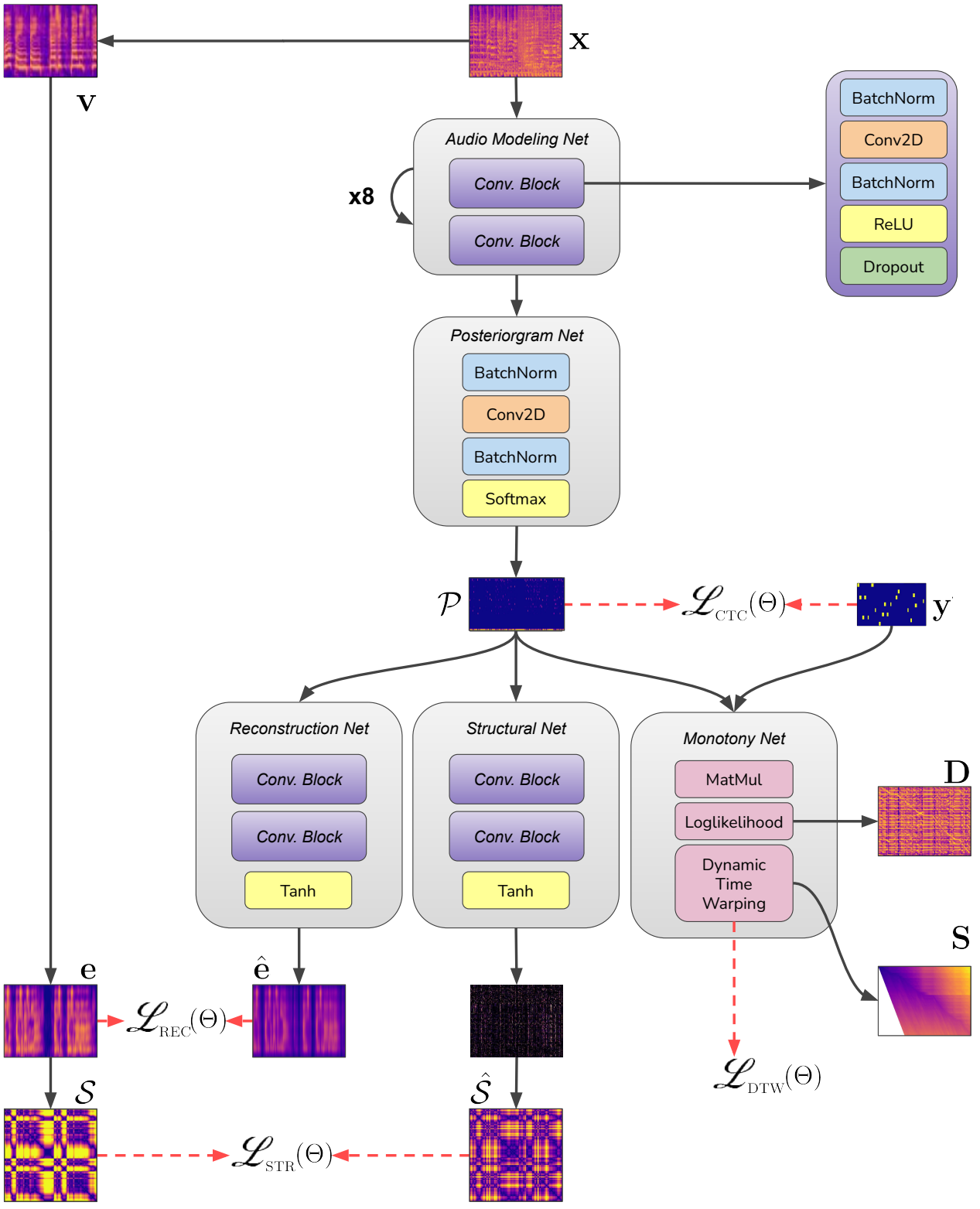


Figure 5.8: ADAGIO enhanced with temporal constraints of spectral reconstruction, structure propagation and guided monotony.

Part III

Outcomes: experiments & applications

Chapter 6

Evaluations of deep voice alignment

*"The history is important because science is a discipline deeply immersed in history. In other words, every **Time** you perform an experiment in science or in medicine, what you're actually doing is you're answering someone, answering a question raised by someone in the past."*

– Siddhartha Mukherjee

This chapter focuses on the evaluation of deep learning-based temporal voice alignment systems, with a particular emphasis on the contributions introduced in this research. Through a series of carefully designed experiments, the effectiveness and accuracy of these proposals are quantified and provide valuable insights into their capabilities and limitations.

As developing and evaluating the models require data featuring speaking and singing voice, the many *voice corpora* used in this manuscript are first presented in section 6.1. The core of the *evaluation procedure*, including alignment retrieval by neural inference and decoding to the classical quantitative metrics, is then detailed in section 6.2. Then, in section 6.3, an *ablation study* is carried out to evaluate the impact of the proposed contributions. Finally, *results* from various voice-to-symbols alignments scenarios are shared, compared to relevant baselines, and commented in section 6.4. The main outcomes obtained in this chapter are summarized in section 6.5.

Dataset	Language(s)	Duration [h]	Voice		Content					
					Transcripts		Alignments			Background music
			Speech	Singing	Text	Phonemes	Words	Phones	Notes	
LibriSpeech	English	1,000	✓		✓					
TIMIT	English	5	✓		✓	✓	✓	✓		
Philos 10	English	0.35	✓		✓		✓			
Att-HACK	French	30	✓		✓	✓				
DALI										
⌞ English	English	220.1		✓	✓	✓	✓		✓	✓
⌞ Playlist50	English	2.6		✓	✓	✓	✓		✓	✓
Hansen	English	0.6		✓	✓		✓			
Jamendo	English	1.2		✓	✓		✓			✓
Chanter RT	French	1.5		✓		✓	✓	✓		

Table 6.1: Overview of all voice corpora used in experiments.

6.1 Voice corpora

This section describes the various speech and singing corpora used for training and evaluating the voice aligners. The origin and role of each dataset in the undertaken studies are specified as well as the nature of their annotation (manual *vs* automatically generated). For the sake of conciseness, Table 6.1 summarizes their main characteristics.

6.1.1 Speech datasets

LibriSpeech

LibriSpeech is a corpus featuring 1,000h of spoken voice from audiobook recordings sampled at 16kHz and read by various speakers. It has been prepared in the context of the LibriVox project and has manually been segmented into small audio excerpts and their associated transcripts (Panayotov et al., 2015). Train, test and development sets for clean and more challenging speech are shared. LibriSpeech has been widely used in the speech recognition literature and, although no text alignment is provided, the audio-text pair allows CTC training.

TIMIT

TIMIT (Zue et al., 1990) is another extremely popular dataset in the speech recognition literature. TIMIT is a multi-speaker dataset of 5 hours of speech, organized in 6,300 sentences (4,000 for training, 1,300 for testing), with available word and phonetic transcripts. Although TIMIT is much smaller than LibriSpeech, respective word- and phone-level manual alignments to the audio are available, making it a relevant test set for both word and phonetic alignments.

Philos 10 (The Problems of Philosophy, B. Russell, Chapter 10)

TIMIT and LibriSpeech feature naturally short or pre-segmented audio excerpts that are necessary to train models due to memory limitation on GPU. Yet, evaluations on these datasets may not be really relevant for real-world applications. Indeed, in practice, one rarely needs to align 5-10s of audio with its text, but rather longer context. This is the problem of (very) long audio alignment, which is known as a limitation for most alignment algorithms and that will be briefly mention in section 7.4. There is thus a need for a long audio-text evaluation dataset.

Instead of artificially creating longer test audio-text pairs by concatenating existing small pieces, it was decided to manually annotate an entire chapter from a publicly available audiobook. Given the extremely time-consuming nature of manual annotations, the annotation is limited to word-level alignments. To report fair evaluations, neither the chosen audiobook nor its reader could be part of LibriSpeech. It has been found that "*The Problems of Philosophy*" by Bertrand Russell (2001) fulfilled these criteria. The 10th chapter of this book, which contains exactly 100 sentences (*i.e.*, 2672 words) and has an audio duration of 21min, has been chosen and collectively aligned. This dataset is referred to as *Philos 10* and is shared with the community[1].

The annotation process, as detailed in the associated publication (Doras et al., 2023), has been done as follows: once the Librivox audio preamble (title and copyright information) was discarded, ADAGIO was used to roughly synchronize the onset of each word with the audio. Given these alignments, temporal markers were extracted and manually corrected in some audio visualization software, *e.g.*, Audacity[2] or Partiels[3]. A spectral display was used to better capture the phonetic evolution and mark the word onsets. Some inconsistencies between audio and text were heard, so the text was modified/corrected accordingly to match the audio perfectly.

Att-HACK

The Att-HACK dataset (Le Moine and Obin, 2020) complements the above-mentioned ones as it exposes (French) expressive speech, *i.e.*, speech acted by professional actors that were asked to sound as natural as possible given some instructions.

Concretely, 20 native French speakers (9 men and 11 women) had to portray four social attitudes – friendly, seductive, dominant and distant – over 100 isolated sentences. To account for the individual variations when it comes to produce a vocal attitude, each utterance has been recorded multiple times (three to five versions), resulting in a total of 30h of expressive speech. All the audio excerpts are provided with their orthographic text transcription.

[1] https://ircam-anasynth.github.io/papers/2023/a-linear-memory-ctc-based-algorithm-for-text-to-voice-alignment-of-very-long-audio-recordings

[2] https://github.com/audacity/audacity

[3] https://forum.ircam.fr/projects/detail/partiels/

6.1.2 Singing datasets

DALI

DALI (**D**ataset of synchronised **A**udio, **L**yr**I**cs and notes)[4] is the first publicly available, large-scale dataset of singing voice with various annotations (Meseguer-Brocal et al., 2018). It contains 5,358 songs, each with note alignments and lyrics alignments at the word, sentence and paragraph levels as well as additional meta information (language, genre, artist, year, etc.). The dataset is coloured in terms of western genres (*e.g.*, pop, rock, rap, etc.) and languages (*e.g.*, English, French, Spanish, Italian, Polish, etc.).

It has been created using the machine learning teacher-student paradigm (Meseguer-Brocal et al., 2020b) from online manual annotations considered as reliable. Its release has permitted significant improvement and many success in various Music Information Retrieval (MIR) tasks involving singing, *e.g.*, (Demirel et al., 2021; Meseguer-Brocal and Peeters, 2020; Renault et al., 2021), since DALI offers sufficient amount of data to train deep models.

Playlist 50

In order to have a unique dataset allowing the evaluations of *both* text and notes alignment for singing voice with polyphonic music, 50 songs (*i.e.*, about 3 hours) and their available annotations were selected to constitute a test set − the so-called *Playlist 50*. It is larger than the classical evaluations sets (see below) and has originally been used to assess very long alignment by Doras et al. (2023). The DALI IDs and annotations are accessible to the community[5].

Hansen (a capella)

Hansen's dataset (Hansen, 2012) is one of the most famous singing datasets for evaluating and benchmarking systems on automatic lyrics transcription and alignment tasks. It is composed of 9 entire pop music songs in English with manual annotations of word onsets and offsets. The version at disposal is the *a cappella* one, hence featuring solo singing voice only.

Jamendo

Jamendo (Stoller et al., 2019) is also a very popular dataset to assess singing transcription and alignment. It features 20 entire music songs (from 10 different Western genres) with manual annotations of word onsets. All songs have instrumental accompaniment allowing to quantify the performances of a model on real-world data and its robustness to background music.

[4] https://github.com/gabolsgabs/DALI/

[5] https://ircam-anasynth.github.io/papers/2023/a-linear-memory-ctc-based-algorithm-for-text-to-voice-alignment-of-very-long-audio-recordings

Chanter RT

The 1.5-hour French Chanter RT dataset of solo singing was recorded to build a singing system based on phonetic concatenation as detailed in (Ardaillon, 2017). As a result, the singer (RT) was asked to sing slowly and hold long vowels to facilitate this synthesis objective. The temporal distribution of the sung phonemes is bi-modal with a lobe around 1s for the vowels and another around 200ms for the other phonemes. It features 36 French phonemes plus a silence token. This dataset will allow testing whether ADAGIO alignments are relevant for concatenative synthesis, which is one of the target applications of this work – see section 7.1.

Section summary – Voice corpora

The diverse voice corpora used in this thesis – for experiments (training phase), evaluations (inference phase) and for voice studies – have been presented. Two of them were proposed in co-authored (Doras et al., 2023) and further shared with the community: (1) *Philos 10* for a manually word-level aligned audiobook chapter; and (2) *Playlist 50* for a larger singing evaluation dataset with word-level notes and lyrics alignments.

6.2 Evaluation procedure

In this section, the complete evaluation procedure is thoroughly detailed from the computed quantitative metrics to implementation details for ADAGIO and baselines.

6.2.1 Assessment metrics

The quantitative evaluation of voice alignment requires the definition of assessment metrics. There already exist standard evaluation procedure and metrics for the alignment task (Cont et al., 2007). In this work, benefiting from the existence of the (in their own words) "transparent, standardized, and straightforward" library `mir_eval`[6] (Raffel et al., 2014), the metrics they proposed will be used and briefly introduced. Note that, in all computations, only the symbol onsets are considered as decision boundaries. Taking end positions into account would penalize a system detecting perfectly each start of utterance without predicting its full duration. This is commonly done in the literature, *e.g.*, in the competitive challenge MIREX[7].

The starting point of any alignment evaluation is the alignment errors between the predicted timestamps and the truth timestamps associated with each event of the symbolic sequence.

[6] https://craffel.github.io/mir_eval/

[7] https://www.music-ir.org/mirex/wiki/2020:Lyrics_Transcription

Average Absolute Error (AAE)

The Average Absolute Error (AAE) reports the *average* over all events of the absolute difference between estimated and reference timestamps. It is the metric most used to evaluate alignment. The lower, the better.

Median Absolute Error (MAE)

The Median Absolute Error (MAE) reports the *median* of the absolute difference between estimated and reference timestamps over all events. In opposition to the AAE, this metric is insensitive to outliers. The lower, the better.

Percentage of Correct Onsets (PCO)

The Percentage of Correct Onsets (PCO) is a metric measuring the percentage of predicted onset timestamps that can be considered correctly aligned. A threshold of 300ms is commonly admitted and chosen by the community for the misaligned/well-aligned binary decision (Cont et al., 2007; Vaglio et al., 2020a), and so do these evaluations. The higher, the better.

Karaoke Perceptual Metric (KPM)

Finally, the Karaoke Perceptual Metric (KPM), introduced as "Perc-PCO" in Mascleff et al. (2021), takes the human synchronicity perception into account and reflects how alignments would be judged synchronous by users in a Karaoke-like scenario. The higher, the better.

The acoustic model is also often evaluated for transcription, see Eq. (2.25), via two metrics.

Character Error Rate (CER)

The Character Error Rate (CER) indicates the percentage (%) of elements (*e.g.*, graphemes, phonemes, notes, etc.) that were incorrectly predicted by a model in comparison to a ground-truth transcription. It is computed with the JiWER toolkit[8]. The lower, the better.

Word Error Rate (WER)

The Word Error Rate (WER) indicates the percentage (%) of groups of elements (*e.g.*, words) that were incorrectly predicted by a model in comparison to a ground-truth transcription. A single error in the group classifies the group as false. This metric is therefore much more drastic than CER. The JiWER toolkit is also used. The lower, the better.

[8] https://github.com/jitsi/jiwer

6.2.2 Implementation details

In this section the practical implementations of the deep neural trainings, alignment retrieval, and baseline comparisons are mentioned.

Data & alphabet

For singing voice, the models are trained on the English part of the DALI dataset which excludes (1) some validation data with unique singers that are *not* part of the training data; and (2) songs from evaluation set *Playlist 50* − none of their different 50 singers are part of the training or validations sets. DALI songs were segmented into small audio excerpts between 10 and 20s (with their transcripts) by making sure not to truncate the last annotated word. For speaking voice, the models are trained on the complete LibriSpeech dataset. The validation set is TIMIT.

Depending on the experiments dfferent alphabets $\mathcal{A}$ will be used. For phonetic alignment, a custom alphabet covering all the phonemes encountered in the associated database is constituted. For text alignment, the augmented Latin alphabet $\mathcal{A}_\ell$ from Eq. (2.1), which includes graphemes, digits and a space token ø with a size of $L = 37$, is considered. For note alignment, a note alphabet $\mathcal{A}_\text{♩}$ is defined. To do so, the F0 annotations in DALI were retrieved and converted into notes. The resulting notes ranged from C_1 to C_7, which is particularly large for the human voice, but a manual inspection of outliers has not been pursued. The alphabet $\mathcal{A}_\text{♩}$ contains all 12 semitones per octave and a silence token for long pauses (*i.e.*, silences longer than 500ms according to DALI annotations), hence a size of $L = 73$.

Training setup

Unless clearly stated otherwise, all experiments rely on the same training procedure.

Training is performed on Graphical Processing Units (GPU) − either with one GeForce GTX 1080 Ti with 11Go or NVIDIA TITAN Xp COLLECTORS EDITION with 12Go − to benefit from their computational power and parallelization strategies. A number of 10 epochs is typically chosen. Each epoch processes the entire database by means of 16-sample (for DALI) or 8-sample (for LibriSpeech) batches. The loss(es) function(s) is(are) minimized with default ADAM optimizer and an initial learning rate set to $\lambda = 1e^{-4}$.

At the end of each epoch, the Average Absolute Error (AAE) is computed over the associated validation set. The learning rate is reduced by a factor 0.8 each time the alignment error has not decreased on the evaluation set for 2 consecutive epochs. The training time for one configuration of ADAGIO on DALI/Librispeech in the given setup ranges between 18h (no constraints) to 29h (all constraints).

Codes are written in Python/Tensorflow and are supported, at the time of writing, in Tensorflow 2.8. The original inspiration was the `CTCModel` implementation as proposed and shared by Soullard et al. (2019).

Alignment retrieval

Remembering the general overview of any voice aligner (see Figure 3.2), an acoustic model $\mathcal{M}_\Theta$ generates an encoding of the audio $\mathbf{x}$ and, second, a dedicated decoding module $\mathcal{D}_\pi$ is used to force-align the target sequence $\mathbf{y}$ to the temporal axis of the encoded representation $\mathcal{M}_\Theta(\mathbf{x})$.

In this context, ADAGIO acts as the acoustic model and generates a CTC posteriorgram $\mathcal{P}$ from the voice recordings, as explained in section 3.5.1, such that $\mathcal{P} = \mathcal{M}_\Theta(\mathbf{x})$. It is the role of the specific CTC decoder, which is the focus of section 3.5.2, to finally retrieve $\hat{\boldsymbol{\pi}}^*$ – the estimated alignment between $\mathbf{x}$ and $\mathbf{y}$.

Baselines & state of the art

The performances of ADAGIO are then compared to those of other alignment systems.

Baselines – In terms of end-to-end neural architectures, CRNN (Vaglio et al., 2020a) and ARNN (Schulze-Forster et al., 2020) have been re-implemented and will be used for polyphonic music evaluation (CRNN) and for phonetic alignment (ARNN) – see section 4.1.1.

State of the art – Results from the latest Music Information Retrieval Evaluation eXchange (MIREX)[9] with lyrics alignment task are also reported as the work of Gao et al. (2021) remains, to the best of the author's knowledge, current state of the art in Automatic Lyrics Alignment (ALA) on classical evaluation datasets. It is based on the system of Gupta et al. (2020), which exploits (1) an acoustic model made of time-dilated neural layers trained with Kaldi (Povey et al., 2011) on the English subset of DALI; (2) an extended pronunciation lexicon addressing the long hold on vowels in singing (Gupta et al., 2018); (3) additional genre-informed modeling of phonemes and silences (Gupta et al., 2020); and (4) a tri-gram word language model created from the available lyrics. It is directly usable on polyphonic music but, as seen, connects several specialized modules, as opposed to end-to-end models. This model is referred to as SOTA.

Section summary – Evaluation procedure

ADAGIO, upon training on a single GPU with publicly available dataset for speech and singing, will be used in inference to estimate voice-to-symbol alignments (predictions) that will be compared to annotations at disposal (ground-truth). The temporal differences will allow to assess the model with four metrics for alignment – *i.e.*, Average Absolute Error (AAE), Median Absolute Error (MAE), Percentage of Correct Onsets (PCO) and Karaoke Perceptual Metric (KPM) –, and two metrics for transcription – *i.e.*, Character Error Rate (CER) and Word Error Rate (WER). Comparisons with existing baselines are performed via the re-implementation of CRNN and ARNN models and latest MIREX challenge results to relate to the state of the art (SOTA) held by Gao et al. (2021).

[9] https://www.music-ir.org/mirex/wiki/2020:Lyrics_Transcription

Configuration	$\mathscr{L}_{\text{CTC}}(\Theta)$	$\mathscr{L}_{\text{REC}}(\Theta)$	$\mathscr{L}_{\text{STR}}(\Theta)$	$\mathscr{L}_{\text{DTW}}(\Theta)$
C	×			
CR	×	×		
CS	×		×	
CD	×			×
CRS	×	×	×	
CRD	×	×		×
CSD	×		×	×
CRSD	×	×	×	×

Table 6.2: Model configurations for ADAGIO according to the loss(es) to minimize.

		Alignment metrics				Transcription	
Dataset	Configuration	AAE [ms]$_{\downarrow}$	MAE [ms]$_{\downarrow}$	PCO [%]$_{\uparrow}$	KPM [%]$_{\uparrow}$	CER [%]$_{\downarrow}$	WER [%]$_{\downarrow}$
Playlist50	C	124.2 (± 29.2)	39.2 (± 3.7)	95.2 (± 1.1)	88.6 (± 1.4)	53.7 (± 1.3)	94.5 (± 1.7)
	CR	96.2 (± 19.2)	39.6 (± 4.7)	96.5 (± 0.8)	87.9 (± 1.3)	49.4 (± 1.6)	91.3 (± 2.2)
	CS	103.4 (± 27.6)	39.7 (± 4.7)	96.4 (± 1.0)	87.9 (± 1.3)	50.1 (± 1.6)	**89.1** (± 1.7)
	CD	98.8 (± 17.3)	38.9 (± 3.0)	96.3 (± 0.8)	**90.4** (± 1.1)	51.3 (± 1.6)	91.2 (± 1.6)
	CRS	102.6 (± 18.5)	39.6 (± 4.3)	95.7 (± 0.9)	87.7 (± 1.3)	51.3 (± 1.5)	91.4 (± 1.5)
	CRD	96.1 (± 13.2)	41.5 (± 3.5)	96.3 (± 0.8)	50.2 (± 1.6)	50.2 (± 1.6)	89.9 (± 1.7)
	CSD	98.5 (± 38.5)	38.5 (± 4.4)	96.6 (± 0.8)	88.5 (± 1.2)	**49.0** (± 1.6)	90.1 (± 2.1)
	CRSD	**93.1** (± 16.4)	**38.0** (± 4.0)	**96.8** (± 0.8)	89.2 (± 1.2)	49.4 (± 1.7)	89.6 (± 2.2)
Philos 10	C	51.2 (± 19.1)	46.2 (± 2.6)	**100.0** (± 0.0)	95.6 (± 2.1)	**15.1** (± 1.1)	**47.9** (± 2.6)
	CRSD	**45.7** (± 20.1)	**39.9** (± 2.1)	**100.0** (± 0.0)	**96.7** (± 1.8)	17.9 (± 2.5)	55.0 (± 2.9)

Table 6.3: Impact of additional temporal constraints (with scaling and supervised on vocals) on voice-to-word alignment. For metrics, ↑ means higher is better, ↓ means lower is better, and 95% confidence intervals on the mean are shown.

6.3 Ablation study

Before comparing the ADAGIO system to reference aligners, its best version must be determined. In view of the proposals introduced in the previous Part II, comparative tests must be conducted on the various versions of ADAGIO based on the possible combinations of additional temporal constraints. This section specifically addresses such ablation studies. The denomination of the model's configurations with respect to their supervision strategy is exposed in Table 6.2.

Impacts of the temporal constraints

The Table 6.3 reports the evaluations of ADAGIO on voice-to-word alignment for its different configurations (these are also depicted on the left side of Figure 6.1 and Figure 6.2).

As singing voice in a musical context is much more challenging to align than clean speech, the observed effects are stronger for singing, hence easier to showcase − that is why, despite similar conclusions for speech and singing, all configurations are shown for *Playlist 50* and only the two extremes (CTC alone *vs* fully constrained CTC) for *Philos 10*.

First, one can see that ***the proposed additional temporal constraints are indeed beneficial to the alignment task***. The very fact to add one supplementary loss already improves the AAE. Interestingly, the MAE remains quite stable among the configuration and around 40ms for both speech and singing voice. The configuration with best potential for *karaoke* application is CD with the best KPM − its alignment would globally be perceived more synchronous in a karaoke-like display. It is also confirmed, as stated in section 5.1, that better performances in *transcription* with CTC do not necessarily result in better *alignment*.

For speech, the AAE is typically half of that observed for singing. This indicates that there are fewer outliers for speech and that these outliers are less severe, which is in line with the absence of a high "corruption" of the voice through the musical accompaniment.

The full combination of constraints, *i.e.*, the final configuration CRSD, again slightly improves the overall alignment accuracy and stands out in AAE, MAE and PCO metrics. Thus, from now on, any mention of ADAGIO will systematically refer to the CRSD configuration.

Impact of scaling the losses

The Figure 6.1 shows, for each configuration, the Average Absolute Error (AAE) measured whether the temporal constraints are scaled or not. One can see the utmost importance of the scaling process to obtain relevant alignment performances.

In the case of the *reconstruction* loss $\mathscr{L}_{\text{REC}}(\Theta)$ − without scaling, the reconstruction dominates the CTC cost such that the model struggles to converge and denegerates into some simple "spectral auto-encoder" that does not perform alignment. This holds true for the structural loss $\mathscr{L}_{\text{STR}}(\Theta)$ with an even stronger degradation as it scales quadratically in audio length.

Interestingly, as the monotony loss $\mathscr{L}_{\text{DTW}}(\Theta)$ does not require any scaling − see Eq. (5.15), the configuration CD is less sensitive to the scaling process. Yet, no scaling leads to a CTC loss dominating DTW, which turns out to have a negative impact on the alignment.

Impact of voice supervision

The Figure 6.2 shows, for each configuration, the Average Absolute Error (AAE) measured whether the temporal constraints are supervised on *mixes* or *vocals* (estimated, as explained in section 5.2.1, via a voice separator based on the work of Choi et al. (2019)).

It is straightforward to conclude that supervision on the vocals is the right option, as the mixes contain music instrumental that does not inform about voice properties and, if considered as ground-truth for the temporal constraints, necessarily confuses the network. Here again, as the monotony loss $\mathscr{L}_{\text{DTW}}(\Theta)$ is independent from any references derived from audio features for its computation − but only requires the available posteriorgrams and transcripts −, the configuration CD is not affected by this study.

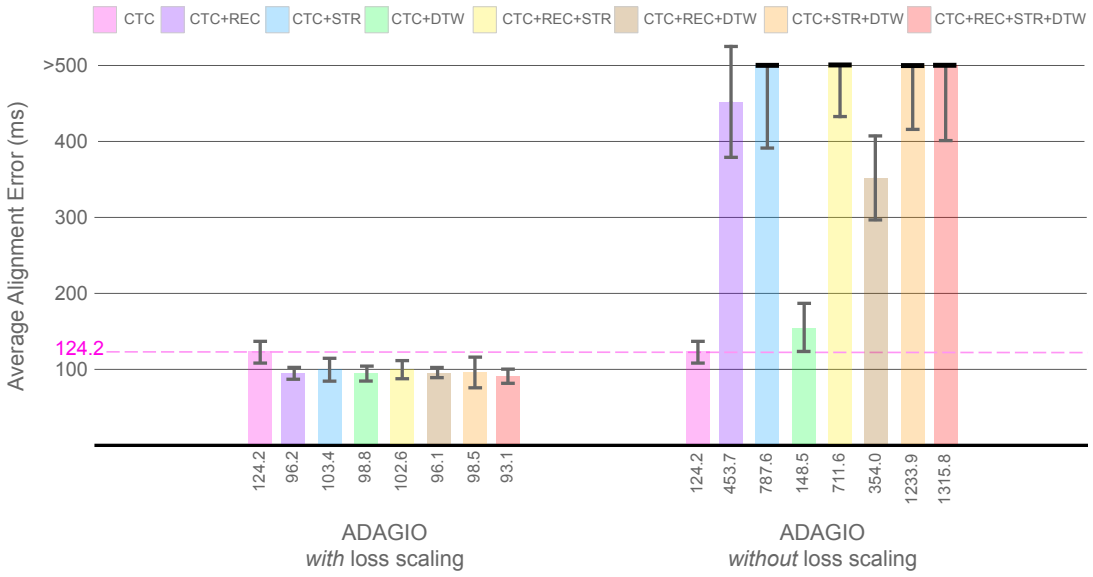


Figure 6.1: Ablation study – impact of scaling the losses. Errors bars correspond to 95% confidence intervals on the mean. AAEs greater than 500ms are masked for readability.

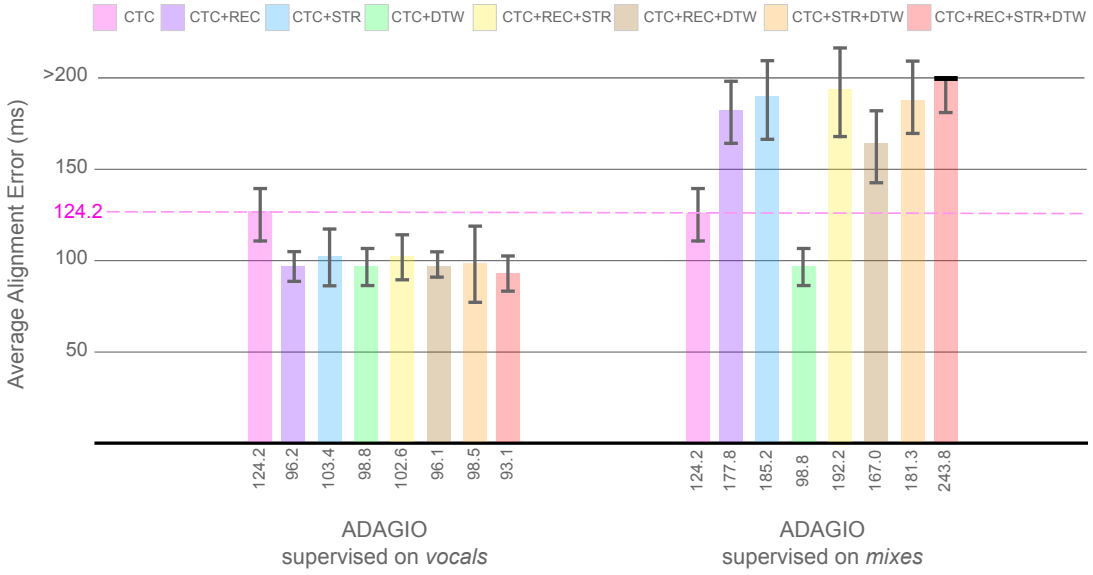


Figure 6.2: Ablation study – impact of the temporal constraints and the nature of their supervision (mixes vs vocals). Errors bars correspond to 95% confidence intervals on the mean. AAEs greater than 200ms are masked for readability.

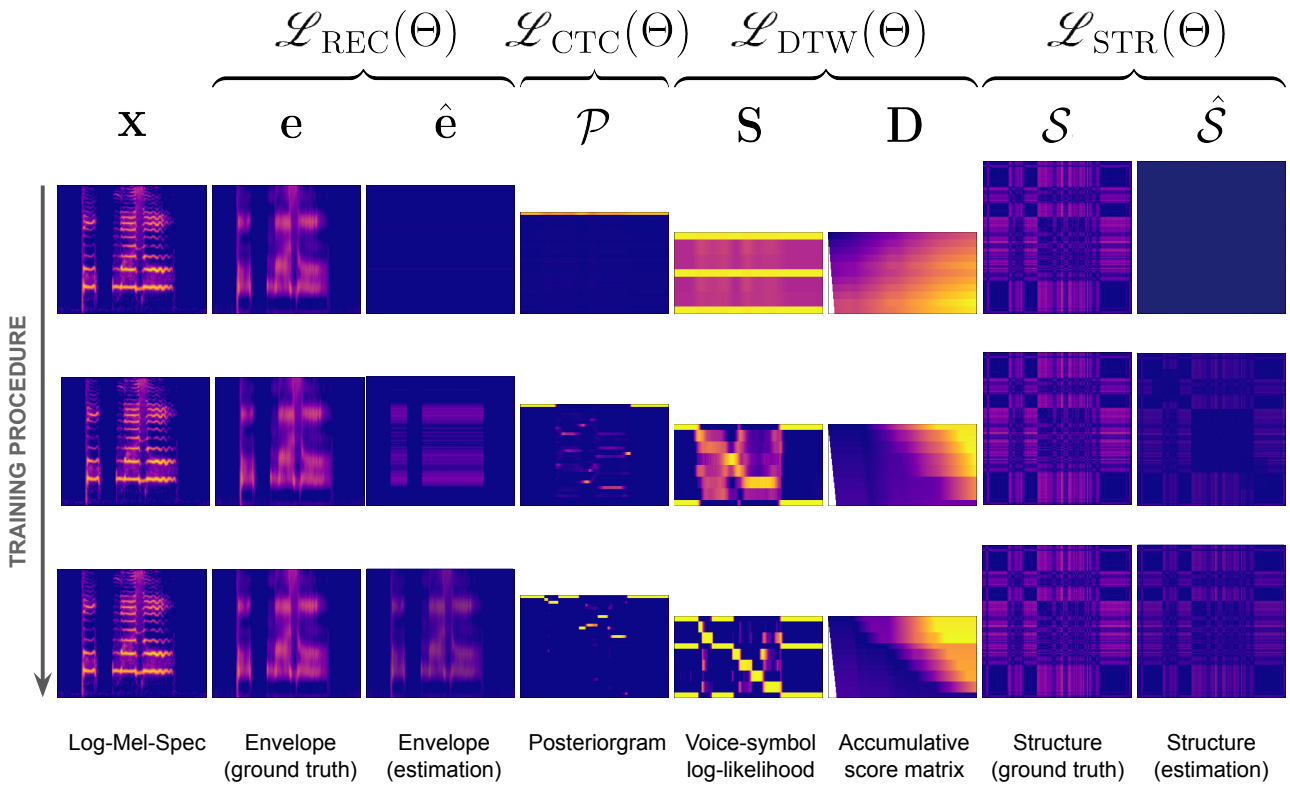


Figure 6.3: Impact of the several temporal constraints during the training phase. The sample used to create the figure was selected from the *validation* set at the end of the first, an intermediate and the final epoch. Example is a clean singing voice dataset. Throughout training, ADAGIO progressively learns not only to recognize the correct phonemes but also to solve the additional constraints leading to better alignments (see evaluations).

Visualizations during training

Finally, for illustrations purposes, the Figure 6.3 shows the evolution of the different losses during training.

At the beginning of the training, no information is captured so that neither recognition nor alignment are actually performed. A little bit at the time, symbols (here, phonemes) are recognized and their selection serves to reconstruct an estimate of the spectral envelope, temporal structure and construct an alignment score. By the end of the training procedure, phonetic content is well recognized and predicted with high temporal relevance so that the additional temporal constraints are also satisfied.

Section summary – Ablation study

Through ablation studies, the relevance of the additional temporal constraints proposed in this thesis was confirmed. The best configuration, in terms of AAE, MAE and PCO, is associated with the full combination of these supplementary objectives and will be systematically implied hereinafter when mentioning ADAGIO. Also, the necessity to scale the various losses and to supervise them on *vocals* – estimated via a source separation algorithm – instead of direct mixes was highlighted.

6.4 Temporal alignment results

This section is dedicated to the evaluation of the optimal version of ADAGIO, according to above-mentioned ablation studies, on various voice databases and its comparison with other algorithms from the literature. On the one hand, voice-to-*text* alignment is evaluated at different granularities – word-level and phone-level text synchronizations. On the other hand, voice-to-*note* alignment is also assessed. These three concrete usages will have their importance in the next Chapter 7 dedicated to the exploitations of such time-symbol mappings.

6.4.1 Voice-to-text alignment

First, the task of synchronizing an audio recording with its text transcript is evaluated. Depending on the chosen text representations – ultimately engraved in the used alphabet $\mathcal{A}$ which can be composed of graphemes or phonemes – two subtasks can be investigated, *i.e.*, alignment of word and alignment of phonemes.

High level alignment – word granularity

The Table 6.4 sets the results for *word*-level alignment out for speech and singing voice. From this table, one can conclude that ADAGIO produces relevant alignment performances that compare favorably to the Convolutional Recurrent Neural Network (CRNN) baseline, even outperforming it on most alignment metrics. The CRNN model, though, remains almost systematically better for pure transcription, which can be explained by its recurrent architecture.

However, SOTA performances measured in Average Absolute Error (AAE) are not reached by ADAGIO on the classical evaluations datasets (Jamendo and Hansen). *Ad hoc* error repartition revealed two strong outliers for 2 out of the 20 songs in Jamendo, and especially an extreme one (>4s) for a song that, interestingly, all MIREX submissions are struggling to align[10]. This naturally degrades the AAE, all the more so given the small number of track samples.

[10] `Pure_Mids_-_The_Leader_`

		Alignment metrics				Transcription	
Dataset	Model	AAE [ms]$_\downarrow$	MAE [ms]$_\downarrow$	PCO [%]$_\uparrow$	KPM [%]$_\uparrow$	CER [%]$_\downarrow$	WER [%]$_\downarrow$
Philos10	ADAGIO	**45.7** (± 20.1)	**39.9** (± 2.1)	**100.0** (± 0.0)	**96.7** (± 1.8)	**17.9** (± 2.5)	55.0 (± 2.9)
	CRNN	79.2 (± 30.2)	56.9 (± 6.9)	**100.0** (± 0.0)	94.9 (± 3.3)	18.9 (± 2.1)	**53.9** (± 3.2)
Playlist50	ADAGIO	**93.1** (± 16.4)	**38.0** (± 4.0)	**96.8** (± 0.8)	**89.2** (± 1.2)	49.4 (± 1.7)	89.6 (± 2.2)
	CRNN	135.4 (± 13.6)	42.5 (± 4.9)	93.5 (± 0.7)	88.4 (± 1.2)	**42.8** (± 2.3)	**81.7** (± 2.2)
Hansen	ADAGIO	115.5 (± 104.1)	47.3 (± 11.9)	97.4 (± 2.6)	84.0 (± 2.9)	**36.8** (± 3.8)	75.6 (± 4.2)
	CRNN	146.5 (± 57.0)	51.3 (± 10.0)	96.2 (± 2.2)	82.2 (± 3.1)	37.4 (± 4.3)	**72.1** (± 4.2)
	SOTA	**86.7** (± 65.6)	**31.6** (± 7.5)	–	–	–	–
Jamendo	ADAGIO	284.8 (± 212.3)	47.7 (± 12.5)	**94.5** (± 2.7)	84.0 (± 3.5)	49.2 (± 4.0)	87.4 (± 4.1)
	CRNN	323.8 (± 89.3)	55.7 (± 4.5)	93.2 (± 2.0)	**84.2** (± 2.2)	**46.4** (± 3.2)	**83.1** (± 3.8)
	SOTA	**217.0** (± 127.8)	**46.1** (± 6.1)	–	–	–	–

Table 6.4: Results on voice-to-*word* alignment. For metrics: ↑ means higher is better, ↓ means lower is better, and 95% confidence intervals on the mean are shown. Results for SOTA are from the latest MIREX challenge (https://www.music-ir.org/mirex/wiki/2020:Automatic_Lyrics-to-Audio_Alignment_Results) and only share AAE and MAE metrics for the two classical evaluations sets Jamendo and Hansen.

Yet, the Median Absolute Error (MAE) values produced by ADAGIO – 40-50ms on all test sets – are in line with the ones reported for SOTA. Therefore, it can be stated that these assessments are very much acceptable remembering that SOTA involves acoustic, language, genre and prononciation models while ADAGIO relies on a simpler, fully data-driven architecture.

Low level alignment – phoneme granularity

The Table 6.5 sets the results for *phoneme*-level alignment out. In this table, the Phoneme Error Rate (PER) metric is defined by analogy with the Character Error Rate (CER).

For speech, TIMIT is used as it also features phonetic alignment. Data are already divided into train and test sets. While all three aligners can precisely align phonemes on clean speech, ADAGIO stands out for its alignment quality together with the other criteria that motivated its construction: absence of recurrent layers (as opposed to CRNN & ARNN), and only processing audio (as opposed to ARNN). Again, CRNN remains a better transcriber. In comparison to the first thesis proposal in (Teytaut and Roebel, 2021) with proper similar configuration, AAE (44.1ms) and MAE (27.3ms) have significantly improved with the ADAGIO architecture.

A second evaluation is done on Chanter RT, which also comes with phonetic alignment ground-truths. Yet, just like the specific nature of this dataset (single pitch point, vowels voluntarily held very long, less than 2h of audio in total), a special alignment procedure is proposed here. Instead of dividing the dataset into train and test sets, it is entirely used for training and test stages. 100 training epochs processing the whole dataset are used.

		Alignment metrics				Transcription
Dataset	Model	AAE [ms]$_\downarrow$	MAE [ms]$_\downarrow$	PCO [%]$_\uparrow$	KPM [%]$_\uparrow$	PER [%]$_\downarrow$
TIMIT	ADAGIO	**18.0** (± 0.7)	14.3 (± 0.5)	**100.0** (± 0.0)	**100.0** (± 0.0)	35.2 (± 4.7)
	CRNN	20.5 (± 1.2)	13.3 (± 0.5)	**100.0** (± 0.0)	**100.0** (± 0.0)	**29.3** (± 3.3)
	ARNN	22.5 (± 1.5)	**12.2** (± 0.6)	**100.0** (± 0.0)	**100.0** (± 0.0)	n/a
Chanter RT (*overfitting*)	ADAGIO	**39.3** (± 1.8)	**28.0** (± 0.9)	**99.3** (± 0.2)	94.7 (± 0.2)	n/a
	CRNN	52.2 (± 1.8)	35.2 (± 3.3)	99.0 (± 0.2)	93.9 (± 0.2)	n/a
	ARNN	44.5 (± 3.4)	34.0 (± 1.8)	**99.3**(± 0.2)	**96.3** (± 0.2)	n/a

Table 6.5: Results on voice-to-*phoneme* alignment. For metrics: ↑ means higher is better, ↓ means lower is better, and 95% confidence intervals on the mean are shown. "n/a": non-applicable when overfitting (texts are known) and for ARNN (phonemes are model inputs).

In doing so, *overfitting* is deliberately intended. Such an uncommon approach in deep learning, as one is usually looking for models with high generalization capabilities and avoid specializations on training data, actually has its interest for aligning small datasets with specific phonetic alphabet (theoretically any custom alphabet) whose small amount of data would prevent the learning of an inference algorithm anyway. However, since the CTC only requires weak labeling for its supervision, the symbol timestamps are new information gained throughout the training procedure. The reported results highlights the capability of ADAGIO to competitively produce phonetic alignments in such a context. The overall process (training and alignment retrieval) took about ∼2.5h. Concrete applications of this strategy will be detailed in section 7.1 and section 7.2.

6.4.2 Voice-to-note alignment

The Table 6.6 sets the results for note alignment out. In this table, the Note Error Rate (NER) metric is defined by analogy with the Word Error Rate (WER) as notes, similar to words for a text, are here represented through a succession of characters, *e.g.*, D4# or 0 for silences, separated by a space.

In the same vein as the previous evaluations, one can see that the recurrent architecture (CRNN) achieves better recognition than the purely the convolutional one (ADAGIO), but the later clearly outperforms the baseline when aligning a singing melody to an audio performance.

This note aligner has been successfully integrated in a musicological study, see section 7.3.3, thus demonstrating the practical feasibility of aligning singing notes with CTC. To the best of the author's knowledge, it is the first time that an end-to-end, CTC-based model addresses note alignment − while note transcription was recently tackled (Weiss and Peeters, 2021). Yet, as all the metrics reported, and notably the Average Absolute Error (AAE) and Median Absolute Error (MAE), remain quite high, these results call for modesty and shall be seen as a first step towards future investigations − see perspectives in section 7.4.

		Alignment metrics				Transcription
Dataset	Model	AAE [ms]$_\downarrow$	MAE [ms]$_\downarrow$	PCO [%]$_\uparrow$	KPM [%]$_\uparrow$	NER [%]$_\downarrow$
Playlist50	ADAGIO	**231.8** (± 73.1)	**105.1** (± 47.3)	**83.6** (± 4.4)	**75.3** (± 4.0)	66.7 (± 5.7)
	CRNN	334.1 (± 100.8)	139.3 (± 39.9)	79.7 (± 3.8)	70.0 (± 3.2)	**59.7** (± 3.5)

Table 6.6: Results on voice-to-*note* alignment task. For metrics: ↑ means higher is better, ↓ means lower is better, and 95% confidence intervals on the mean are shown.

In terms of cross-modality comparisons, two statements can be made. First, regarding *recognition*, it appears than musical transcription is easier than textual transcription − although reported performances are far from being exploitable for, *e.g.*, automatic melody transcription. Second, with an AAE and MAE more than twice as large for note alignment than for word alignment, it is clear that the voice-to-note (V2N) alignment problem is more challenging than its voice-to-word (V2W) counterpart.

Three reasons come to mind: (1) for V2N, *each* note is individually aligned (with background music, as opposed to phonetic alignment) whereas for V2W, *several characters* are in the end merged into a single timestamp, which may forgive some local mistakes; (2) the F0 ground-truth in DALI, which is used to obtain the note sequences, is not as reliable as lyrics annotation[11]; and (3) the musical background is in the *same* modality as the symbolic sequence to uncover in the case of V2N so that the network must learn to distinguish two types of musical information (*i.e.*, singing voice pitches and chords), often coherently and harmonically mixed, which may be more complex task than "fully" ignoring the musical context to focus exclusively on voice production mechanism.

For the sake of illustration, the Figure 6.4 displays an example of inference with this model.

6.4.3 Robustness to transcription errors

In practice, the symbolic transcriptions associated with an audio are error prone and can typically contain mistakes. For instance, a text composed of words can contain typos − *e.g.*, one or several characters might have been added, removed, or altered (wrongly spelled for instance), or entire words could be inserted or deleted. In singing, lyrics also carry their own level of uncertainties as onomatopoeia, words or even lines or paragraphs (*e.g.*, spoken interlude in a song) that can actually be heard in the recording might be missing in the transcript. This holds true for music scores which, depending on the context and the singer's actual interpretation of the score, might not represent all notes really sung in a given performance.

[11]This is even explicitly stated on the official DALI webpage: https://github.com/gabolsgabs/DALI (accessed 04/04/2023)

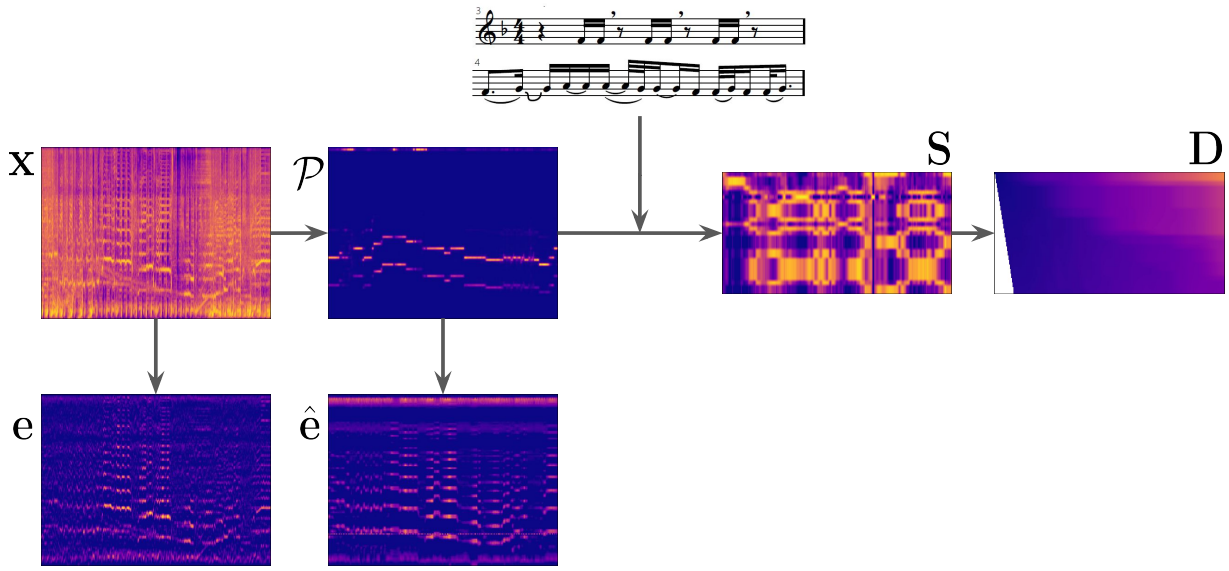


Figure 6.4: An example of inference for voice-to-note alignment with ADAGIO. Note that in this context **e** represents an excitation signal (all but spectral envelope). The posteriorgram, temporally constrained during training, predicts per-frame probability of note occurrences.

In this section, the robustness of ADAGIO to these kinds of transcription errors is evaluated for voice-to-word alignment. To this aim, the alignment accuracy is measured on *Philos 10* and *Playlist 50* with modified transcripts. Similar studies were presented in the publication (Doras et al., 2023) yet, although close conclusions are made, these are distinct evaluations.

Effect of replaced characters

Figure 6.5 first shows the impact of character substitution. For these investigations, p% of the characters are randomly replaced in both *Philos 10* and *Playlist 50*. The space characters ø are not modified so that the number of words is not altered.

The new characters are sampled uniformly at random in the letters (a, ..., z). The audio posteriorgram $\mathcal{P}$ is not altered by these transcriptions as the acoustic model $\mathcal{M}_\Theta$ exclusively processes the audio. The difference is that the decoding module $\mathcal{D}_\pi$ here tries to force-align $\mathcal{P}$ to the altered transcript.

For *speech* – one can see that the alignment error remains remarkably stable around the baseline (*i.e.*, true transcript) up to 50% of replaced characters. The error even remains under a 200ms threshold up to 80% of substitutions, but dramatically rises after that. The system is thus able to compensate for transcription errors, as long as it obtains enough correct anchorage characters, which is guaranteed by the space token ø.

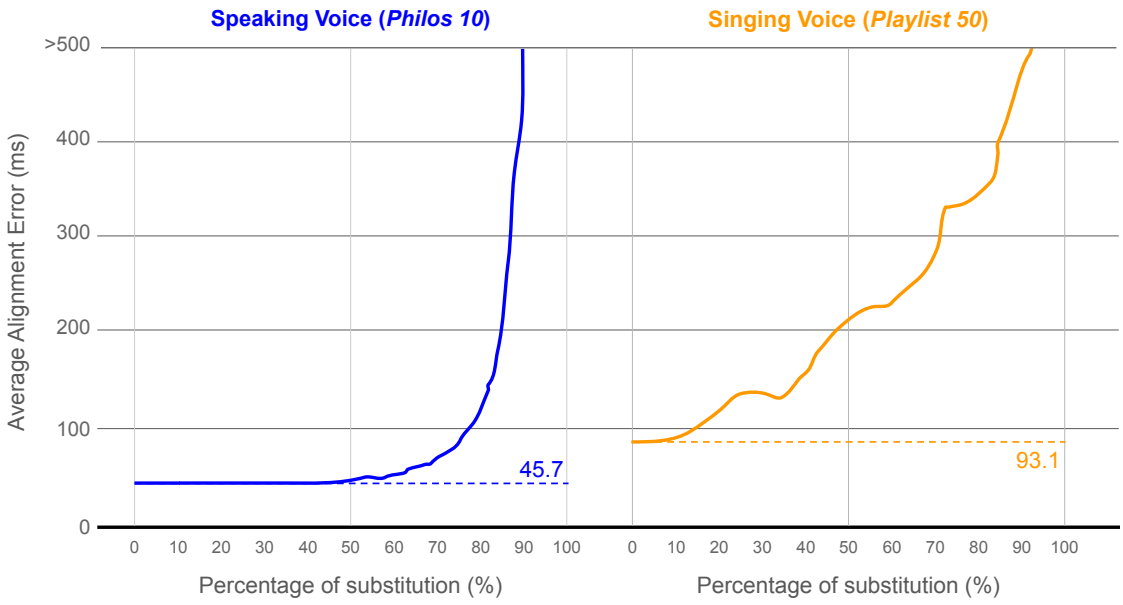


Figure 6.5: Average Absolute Error (AAE) for different percentages of substituted characters for speaking and singing voices with ADAGIO.

For *singing* – the error increases much faster than for speech, even for a small percentage of replaced characters. Due to the presence of musical accompaniment, the posteriorgram exhibits less clear-cut character probabilities, which causes the correct anchor characters to be more easily confused with wrong characters and the optimal alignment path to diverge faster during the decoding step.

Effect of added/removed characters

Figure 6.6 then shows the impact of character insertion and deletion. For these investigations, $p\%$ of the characters are randomly added to or removed from both *Philos 10* and *Playlist 50*. The space characters ø are not modified so that the number of words is not altered. When p decreases towards -100%, it becomes likely that several characters will be removed per word – it is thus ensured that at least one character remains for each word. The alignment path is estimated between the posteriorgram and the new transcript.

For *speech* – the alignment accuracy remain extremely stable around the reference (true transcript) up to 50% of the added/removed characters. The error increases steadily when adding extra characters, as expected. More surprisingly, alignment becomes a bit better when removing 0–30% of the characters. An interpretation for this effect is that removing some characters could be beneficial for the CTC decoding should the model fail to recognize them properly. When more than 50% of the characters are removed, the error increases again.

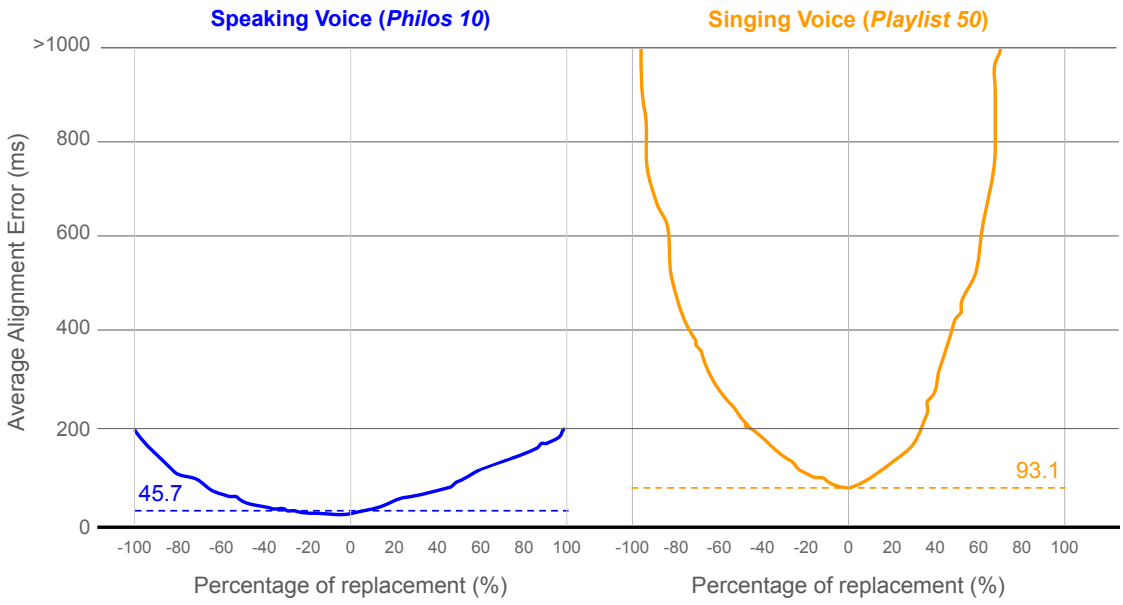


Figure 6.6: Average Absolute Error (AAE) for different percentages of character insertion and deletion for speaking and singing voices with ADAGIO.

Interestingly, the Average Absolute Error (AAE) always remains under 200ms, even after each word contains only one single character. This indicates and confirms that the space character ø is a very powerful anchor, and plays a crucial role in word-level speech alignment.

For *singing* – once again, the model is way more sensitive to these alterations. This can be similarly explained by the fact that anchor characters are more ambiguous for singing than for speech, which makes the optimal path decoding more difficult and more subject to divergence.

Effect of added/removed words

Figure 6.7 finally shows the impact of *word* insertion and deletion. For these investigations, $p\%$ of the words are randomly added to or removed from *Philos 10* and *Playlist 50*. For each added word, a length ℓ is randomly drawn from a normal distribution with a mean of six characters, and then ℓ characters are sampled uniformly at random in the letters (a, ..., z). In these evaluations, the AAE is computed only on the original words, for which true timestamps exist.

For *speech* – the alignment error remains under 200ms error even when every other word is removed or added. This indicates that the system is very robust to noise as long as more than half of the true words remain untouched. However, at some point, there are not enough words to keep the alignment stable, dramatically increasing the AAE.

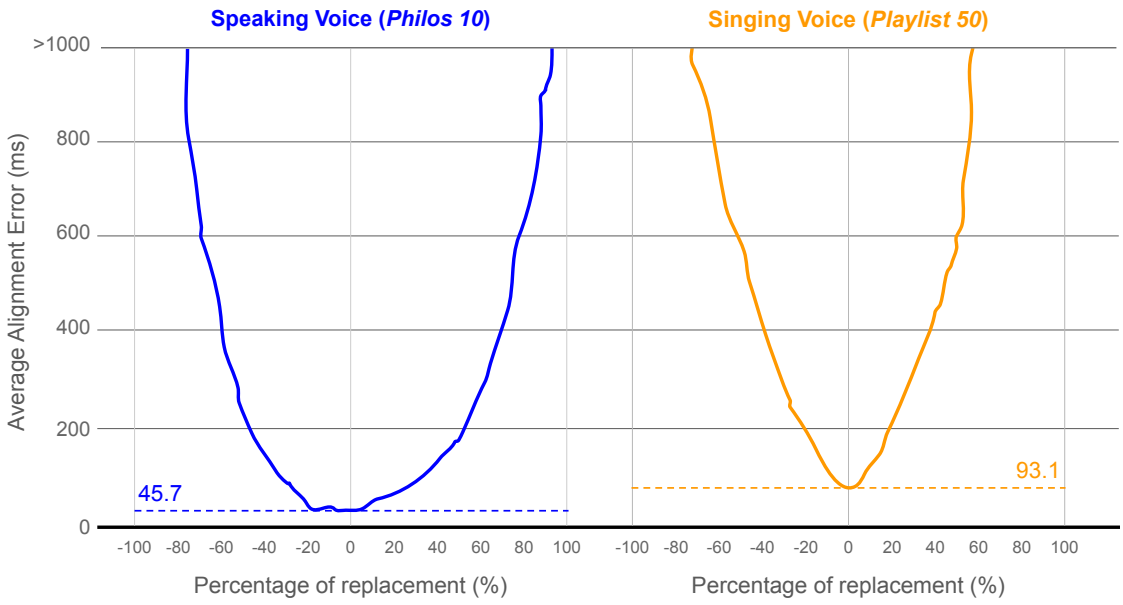


Figure 6.7: Average Absolute Error (AAE) for different percentages of word insertion and deletion for speaking and singing voices with ADAGIO.

For *singing* – similarly to the previous experiments, one can see that singing voice alignment is much more sensitive to a full word addition or deletion. This can be interpreted by the fact that the posteriorgram does not exhibit sufficiently salient probabilities for correcting characters in order to provide steady anchorage compensating for transcription errors.

Section summary – Temporal alignment results

According to various experiments with data with reference alignments, ADAGIO was demonstrated to produce relevant voice-to-symbol alignments in line with recent baselines. First, text alignment has been studied at the word level (by classical inference) and at the phoneme level (by overfitting strategy) and has shown relevant performances with a Median Absolute Error (MAE) below 50ms for both types of voice. The robustness to imperfect transcripts was also investigated and revealed that the system can keep aligning as long as a minimal amount of correct characters serve as anchors for the decoding. Although inferior to the state of the art on lyrics alignment, ADAGIO stands out by its simplicity and ability to be easily transferred from one domain to another thanks to its end-to-end architecture. Second, albeit to a lesser extent, note alignment with ADAGIO was also tackled as a first step towards further development.

6.5 Evaluations of deep voice alignment in a nutshell

Chapter summary – Evaluations of deep voice alignment

In this chapter, I quantitatively showed that my proposed contributions – ADAGIO and its enhancement via temporal constraints – were well-suited for the synchronization between voice and symbols. Based on various *voice corpora*, two of which introduced in this research as evaluation sets and shared with the community, I conducted key experiments to evaluate my acoustic model ADAGIO. First, through an *ablation study*, I demonstrated that all additional temporal constraints ultimately have a positive effect on the alignment accuracy. Then, I compared the best version of ADAGIO, fully time-constrained, to recent baselines and state of the art for text and note alignments. I highlighted the various associated *temporal alignment results* to gather insights from these assessments. Concretely, ADAGIO can perform word-level and phone-level alignment with a median error below 50ms for both speaking and singing voices. Although state-of-the-art performances are not reached for lyrics alignment, ADAGIO stands out by an end-to-end architecture, allowing great flexibility. Finally, note alignment was shown to be feasible and promising but, at the time of writing, still requires further investigations.

Chapter 7

Applications and collaborations

"Hier encore, j'avais vingt ans, je gaspillais le ***Temps*** *en croyant l'arrêter, et pour le retenir, même le devancer, je n'ai fait que courir, et me suis essoufflé."*

Hier Encore (1964)
– Charles Aznavour

The whole manuscript, up to this point, has consisted in the exhaustive presentation of the alignment task between voice and symbolic data as well as its practical implementation through AGADIO – an acoustic neural model reinforced by temporal constraints – whose relevance has been validated by quantitative evaluations. It is now *time* for practice. This chapter, therefore, presents various concrete applications of the ADAGIO algorithm. **These results are from *collaborations* with other researchers and must be seen as proofs of concept of the ADAGIO aligner, rather than original research work done solely by the author.**

In section 7.1, the task of *singing voice synthesis* based on the concatenation of a phonetically aligned singing dataset is addressed. In the following section 7.2, speech phonetic alignment is used to identify some *temporal production strategies* involved in the expression of social attitudes. Continuing the study of vocal expressivity, the next section 7.3 focuses on the *musicological characterization* of singing style, which is at the heart of artistic choices in vocal performances, through syllabic and note alignments. Finally, some *ongoing research and future works* are mentioned in section 7.4. A summary of the chapter is proposed in section 7.5.

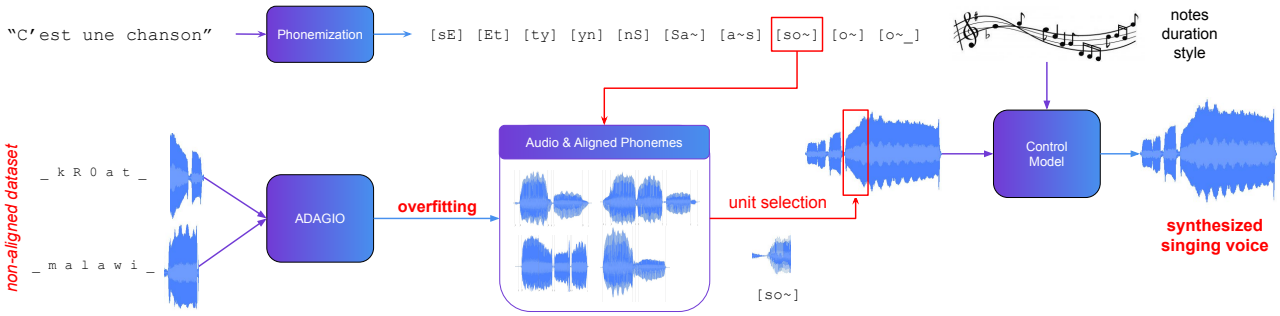


Figure 7.1: Global view of a concatenative synthesis system. To generate an utterance, (di)phonetic units are selected from a pre-aligned database; units are successively concatenated and transformed according to some target parameters. Phonetic alignments of the reference dataset are thus required beforehand. Highly inspired from Figure 3.1 of Ardaillon (2017).

7.1 Concatenative singing synthesis (ISiS)

As first concrete application, temporal voice alignment is used as a means to enrich a singing voice synthesis algorithm based on a concatenative approach which, by nature, requires precise knowledge of the phonetic regions in a reference pre-aligned database. The core of the singing synthesizer – Ircam Singing Synthesis (ISiS) [1] – has been developed in Ardaillon (2017)'s thesis and has not been modified in this work. The application of ADAGIO is an upstream contribution, independent from the synthesis motor *per se* yet necessary to its functioning.

Context

An overview of the synthesis strategy can be found on Figure 7.1 and is briefly summarized. An user asks to synthesize a *target utterance* (*e.g.*, "C'est une chanson") associated with given *control parameters* (*e.g.*, note pitches and durations, singing style, etc.). The desired text is firstly turned into a phonetic sequence (*e.g.*, "s E t y n S a s o _") with a dedicated algorithm. The resulting phonemes are grouped into *diphones*, *e.g.*, succession of two phonemes as [sE] [Et]. The theory of diphones stipulates that transitions between diphones are smoother than between raw phonemes. Then, relevant sound units are selected from an existing singing database that (1) covers all diphones of the language; and (2) *has been temporally aligned*. For each diphone, the unit corresponding best to its surrounding context is chosen. Units are successively concatenated resulting in a first version of the synthesis. The final step consists in transforming the pitches, durations and transitions between the units with signal processing tools to fit some control parameters, *e.g.*, music score or singing style (Ardaillon et al., 2016).

[1] https://forum.ircam.fr/projects/detail/isis/

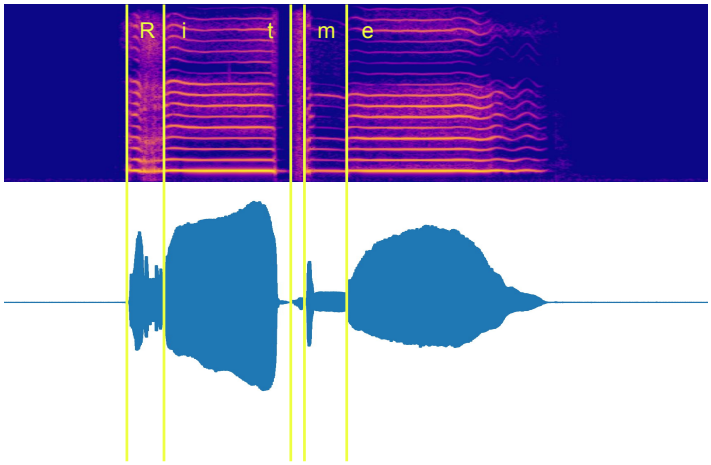


Figure 7.2: Example of a sound sample found in Chanter RT, a reference corpus for ISiS, with phonetic alignment of the French locution "Rythmé" (*rhythmic*) phonemized as "R i t m e".

Unit selection, as a fundamental part of this system, is heavily dependent on the phonetic alignment quality for the reference corpus of sounds. At the time of ISiS creation, an industrial partner with proprietary algorithms was in charge of aligning the corpus. In order to diverse the potential of ISiS, *e.g.*, adding new voices and languages, a voice-to-phoneme aligner is necessary. (The Ircam Analysis-Synthesis team voice aligner, "ircamAlign" (Lanchantin et al., 2008), is not adapted to ISiS corpora as its statistical models were trained for speech.) This is where ADAGIO comes into play. In the context of **Ircam's UPI 2022 "ISiS Voices"**, new singers were recorded for ISiS. The ground-truth phonetic sequences are known but are not aligned – which opens an interesting opportunity for ADAGIO.

Alignment strategy

Currently, only French language is supported in ISiS but extension to other languages is highly desirable. The recorded datasets are small (less than 2h of audio) and composed solely of short audio excerpts (around 3s) in which vowels are purposely held long to facilitate "diphonic" unit selection. An example is depicted in Figure 7.2. Theoretically, any phonetic alphabet could be used as each language would rely on different phonemes. The International Phonetic Alphabet (IPA) being quite large and not fully covered in all languages does not seem as interesting as having a specific and dedicated phonetic catalog per language.

All in all, training ADAGIO as an acoustic model capable of inferring phonetic posteriorgrams for *all* languages seems vain, both due to the lack of available data and potentially infinite diversity of phonetic representations encountered in practice.

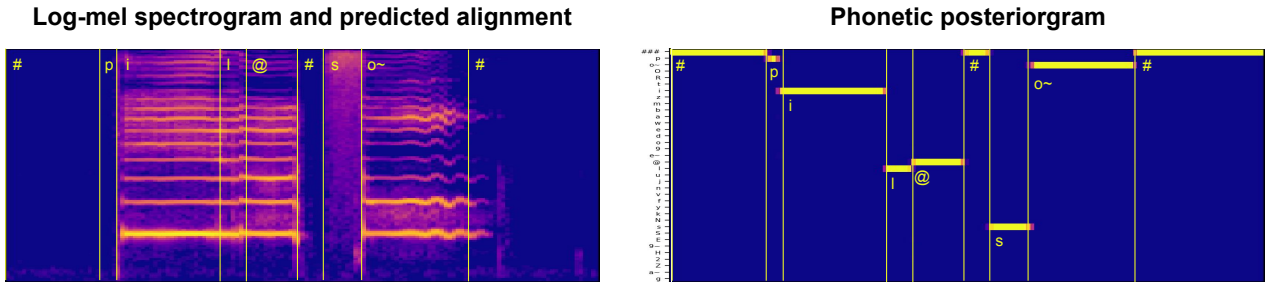


Figure 7.3: Phonetic alignment produced on a newly recorded voice with ADAGIO overfitting.

For these reasons, an *overfitting* approach was chosen. The training of ADAGIO does not aim at achieving a generic acoustic model (*e.g.*, capable of generalization) but at specializing on the target database. Inference, indeed, is made on the same data as during training. The predicted posteriorgrams are then used to predict a forced phonetic alignment.

While everything is usually done to prevent a system from overfitting, temporal information has been gained thanks to this procedure: phonetic sequences, which were not aligned, now are associated with timestamps.

Results

The relevance of this procedure was confirmed in the evaluations conducted in section 6.4 for the Chanter RT dataset − which is the default sound corpus used in ISiS (Ardaillon et al., 2016). Given all of these considerations, a new voice for ISiS has been recorded and aligned by ADAGIO. A result is shown in Figure 7.3. Recording of other singers are on the way. Again, one of the key potential offered by ADAGIO through such usage − intrinsically due to the generic framework of the Connectionist Temporal Classification (CTC) −, is the ability to align any phonetic alphabet in the future, thus extending ISiS to other languages.

Audio examples of voice synthesis with this procedure are accessible online[2].

Section summary − Concatenative singing synthesis

Ircam Singing Synthesis (ISiS) relies on a concatenative strategy to generate sung utterances from pre-aligned dataset at the phoneme level. ADAGIO has been successively applied to align a new French voice thanks to an *overfitting* strategy. This method was validated in previous evaluations for a reference dataset for ISiS with gold ground truth. In the future, ADAGIO will allow the integration of potentially any (phonetic) alphabet in this system and thus, notably, its extension and usage beyond French.

[2] https://www.youtube.com/live/O5RWUl_vZ9M?si=cm2w1qmvC3hjP5ti&t=2128 (35:27)

7.2 Temporal production strategies of vocal attitudes

Depending on the outcome humans expect from an interaction and the nature of the relationship between the individuals − *e.g.*, strangers meeting for the first time, longstanding friends, hierarchy in job interview, etc. −, voice and its expressivity are modulated to convey different social attitudes. The temporality involved in these communicative strategies has not been a core focus of previous research. Benefiting from phonetic alignment, this collaborative work has precisely aimed at uncovering some of the mechanisms at the heart of such expressive speech based on these temporal aspects. This section showcases a summary of associated findings − further details can be found in the complete publication (Salais et al., 2022).

Context

Human interactions rely on communicating social attitudes to other individuals (McAleer et al., 2014). These attitudes, that are distinct from emotions, indicate an individual's social intentions such as being friendly or dominant (Wichmann, 2000). Despite their fundamental role in spoken interactions, the way these attitudes are conveyed through vocal communication has been barely explored. This collaborative study, therefore, has aimed at developing an acoustic evaluation based on anatomic considerations derived from phoneme-to-audio alignment, to uncover the production strategies at play when speakers communicate social attitudes.

To do this, an analysis procedure has been conducted on a 20-speaker subset of the Att-HACK dataset (Le Moine and Obin, 2020), introduced in section 6.1, which features acted expressive speech in four attitudes − dominant, friendly, seductive and distant. The texts were converted into phoneme sequences with the phonemizer of Bernard and Titeux (2021).

Alignment strategy

In order to investigate both articulatory and phonetic information in these vocal strategies, speech signals needed to be segmented so that each phonetic region was clearly identifiable.

In this specific use case, one is not interested in developing a system for inference on new, unseen data, but rather to fully align a fixed dataset of paired audios and unaligned phonetic transcripts. This is possible, once again, with the CTC framework in an *overfitting* scenario − instead of the common seek for *generalization* properties. Therefore, the acoustic model is trained to overfit on the entire Att-HACK subset, and forced phonetic alignments are derived from the posteriorgrams.

Exploitations

The temporal alignment information was then levegared to study both articulatory strategies and phonetic structure.

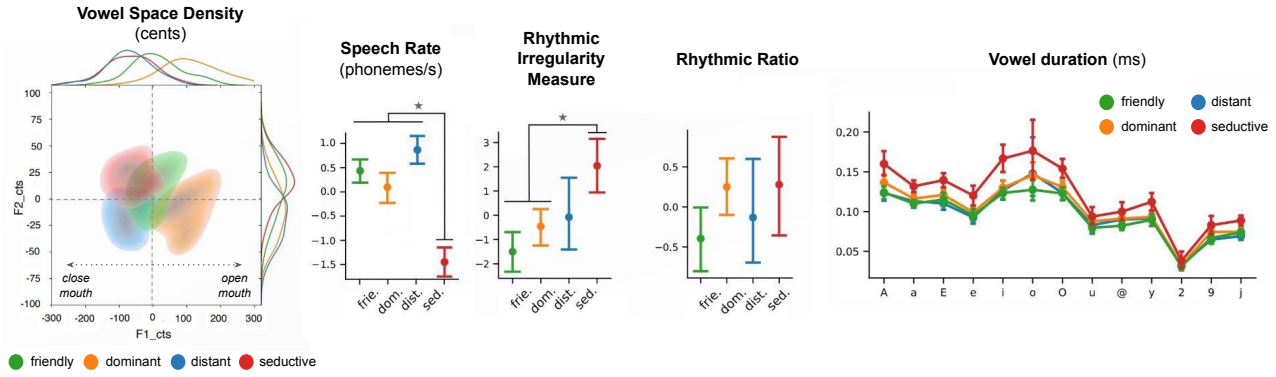


Figure 7.4: Computation and analysis of vocal tract actuation and phonetic structure extracted from voice temporal alignments for friendly, dominant, distant and seductive vocal attitudes. '⋆': statistically significant difference ($p<0.05$). Error bars represent 95% confidence intervals on the mean.

Regarding the *articulatory strategies* – as switching between articulatory modes corresponds to transitions between formants (Pisanski et al., 2022), the Vowel Space Density (VSD) was examined as proposed by Story and Bunton (2017). It represents the space formed by the two formants F_1 and F_2 for all vowels, whose time frames can be known thanks to the alignment, and by only keeping the highest density regions per attitude.

Regarding the *phonetic structure* – revealing the speech prosodic stress and rhythm –, the phonetic timestamps were used to compute (1) the mean phoneme duration, (2) the Speech Rate (SR) as the mean numbers of phoneme per second in each utterance, (3) the Rhythmic Irregularity Measure (RIM) as the mean duration difference between all segments in a sentence, and (4) the Rhythm Ratio (RR) as the mean duration difference between contiguous speech segments. For each attitude, the mean value of these descriptors (over all sentences) was computed upon speaker normalization and zero-centering.

Results

The obtained results are depicted in Figure 7.4.

In terms of *vocal tract articulation*, the Vowel Space Density shows that Distance has a significantly lower F_1 in comparison to Dominance that has a high F_1. This suggests that speakers wanting to establish distance from their audience *shorten* their vocal tract using mostly *closed* mouth, as if being understood by other(s) was not important. Conversely, people desired to be dominant rather *elongate* their vocal tract and *open* their mouth, as a necessity to be understood. More broadly, clusters for each attitude – albeit associated with mostly marginal differences – are identifiable, indeed pointing to an articulatory dimension in speech intention communication, which is a novel perspective.

In terms of *phonetic structure*, significant inter-attitude variations are found for the Speech Rate (SR) and the Rhythmic Irregularity Measure (RIM) but not for Rhythm Ratio (RR). This indicates that attitudes impact *global* rhythmic communication schemes but not *local* ones.

In view of these results, it appears that humans express Seductiveness with a specific temporality, clearly distinct from the other three attitudes studied. As a result, a prototype of "seductiveness" featuring these *temporal* strategies can be highlighted. Specifically, seductive utterances are significantly associated with small speech rate (in line with longer vowel duration) and high RIM, such that they are produced slowly and with irregular rhythm, as if speakers were purposely taking the time to express themselves and hint at their intentions.

Section summary – Temporal production strategies of vocal attitudes

Phonetic alignment has permitted a deep look into the temporal aspects involved when communicating social intentions. It has been shown that French speakers use shared production strategies to express vocal attitudes such as friendliness, dominance, distance or seductiveness. From a phonetic structure perspective, the later attitude (*seductiveness*) is significantly different from the others as associated with slower speech rate and more irregular rhythmics. Besides, and to the best of the authors' knowledge, this study has revealed for the first time that social attitudes are also conveyed via articulatory modes. Insights from significant observed effects are that people might shorten their vocal tract to impose *distance* with their interlocutor and, conversely, elongate it to be *dominant*.

7.3 A musicological pipeline for singing style analysis

In the continuation of the analysis of voice expressivity, this section focuses on the production strategies at play in singing performances, referred to as *singing voice style*.

The study of singing voice style is of great interest both for expressive vocal synthesis and for the musicological analysis of vocal performances and incites to a fruitful convergence between signal processing and musicology. A previous research in that direction (Ardaillon et al., 2016) has precisely been integrated in Ircam Singing Synthesis (ISiS) for expressive control of voice rendering, see section 7.1, and might call for more findings. However, for musicologists, these studies often come up against the absence of semi-automatic analysis tools for voices recorded in a musical context, imposing long and tedious manual annotation work.

This collaboration introduces a complete processing chain in support of musicological analysis, notably providing musicologists with powerful tools for the automatic analysis of singing voices. ADAGIO is a core element of this pipeline, as it is in charge of producing both voice-to-*syllable* and voice-to-*notes* alignments necessary for the fine-grained exploration of artistic choices made by a singer during their performances. This section is based on the collaborative publication (Teytaut et al., 2023) and was made possible by the **ARS** project (https://ars.ircam.fr).

7.3.1 Musicological context

The singing voice, as a vector of communication, appropriates many of the codes of the spoken voice to convey emotions or social attitudes (Lacheret-Dujour and Beaugendre, 1999). As a result, in the 2010s, some musicologists initiated research in vocal performance analysis (Chabot-Canet, 2008; Lacasse, 2010) based on, *e.g.*, paralinguistics, phonostylistics or psycholinguistics (Fónagy, 1983; Léon, 1993; Poyatos, 2002).

Singers can exploit a rich *palette* of singing techniques to shape their interpretations. These techniques can either relate to *general* musical conventions or "trends" – *e.g.*, belting in musicals (Henrich Bernardoni, 2020) or singing format in classical singing (Sundberg, 1974) – or rather a *personal* style via particular vibrato, timbre, phrasing, intonation and more.

The very essence of musical interpretation implies that there may possibly be a large, if not infinite, number of performances corresponding to one piece of music or song. However, an artistic identity necessarily means a certain amount of coherence between the performances – which is ultimately and fortunately fixed, to a certain extent, in some reference recording. Such recordings allow to study, in the most neutral, objective and reproducible way possible, the production strategies and the *palette* of effects deliberately chosen by the singer to be "engraved" in the studio – *i.e.*, the singer's very own *singing voice style* (Chabot-Canet, 2020a).

For years, musicologists interested in singing performance were facing the absence of tools for automatic acoustic analyses of voices recorded in a musical context. Consequently, they had to manually transcribe and synchronize the audio by ear or via visual spectral representations. While expert listening of musicologists remains essential for data supervision and analyses, a purely manual approach has its own limits. It is very time-consuming, imposing one to limit oneself to small corpora, can be error-prone and, in some cases, overly subjective. Thus, relying on automatic systems for transcription (Gao et al., 2022) and alignment (Doras et al., 2023) may lead to a considerable gain in time, and help in setting a common base for the musicological community.

This is the context in which a new musicological pipeline is introduced, with the aim to highly simplify tedious steps previously carried out by hand and opens new perspectives through automation.

7.3.2 Pipeline introduction

This section presents the proposed pipeline for the musicological analysis of singing voice style. As shown in Figure 7.5, it is composed of four main categories, with independent modules (hence a flexible worflow) that are connected, namely *voice characterization*, *musicological expertise*, *voice alignment*, and *musicological exploitations*, that are further detailed.

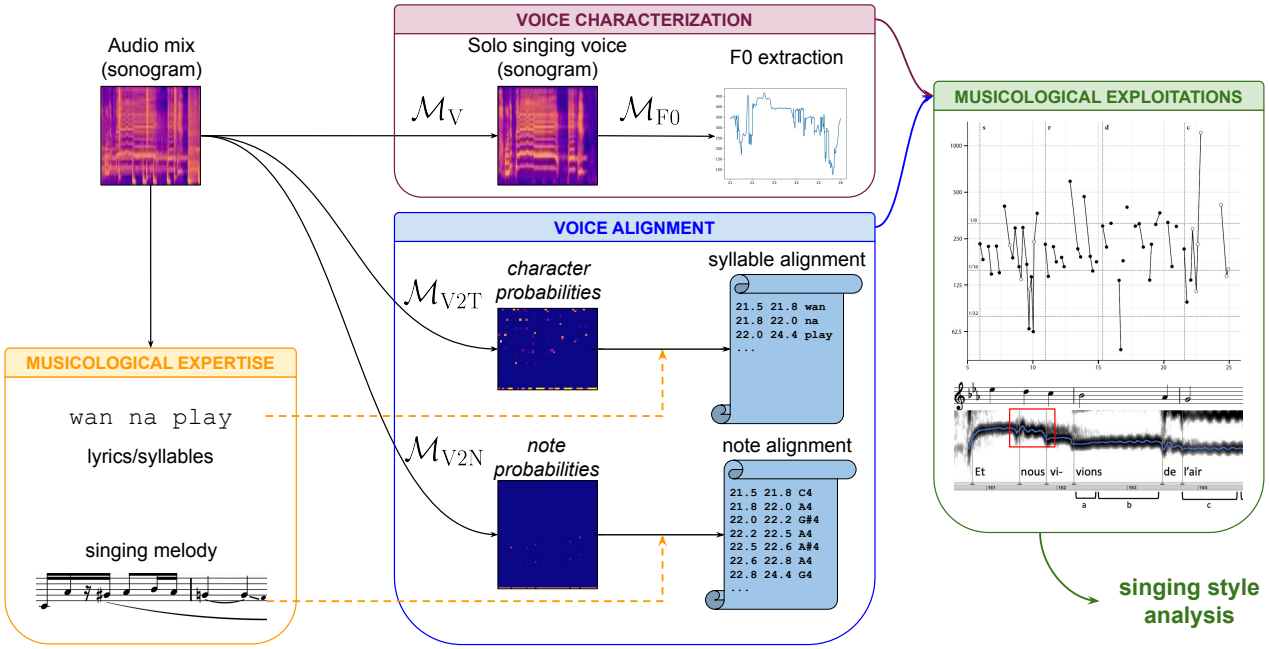


Figure 7.5: Overview of the complete analysis pipeline involving musicological expertise, deep learning models for the automation of voice characterization and alignment in order to help musicologists studying singing voice style.

Voice characterization

(This research has been conducted by co-author Axel Roebel. It is detailed as the voice separator was used to estimate the vocals supervising the temporal constraints in section 5.2.)

The core of any singing or singer analysis system is the voice itself. As musicologists are interested in commercial recordings, the background music must be removed as it is not compatible with parameter estimation algorithms and can hinder a precise description of the singing intonation. The first step, therefore, is to isolate the vocals. To do so, Choi et al. (2019)'s model was re-implemented as it achieves state-of-the-art singing voice extraction quality with a comparatively small number of parameters. This model, denoted $\mathcal{M}_{\mathrm{V}}$ in Figure 7.5, has been trained using the publicly available MUSDB18 (Rafii et al., 2017) and CCMixter (Liutkus et al., 2015) datasets, and a collection of internal data featuring solo singing voices and instrumental music with notably instruments not well covered in the public datasets. During training, the voice and music samples were randomly mixed and pitch-shifted following (Cohen-Hadria et al., 2019; Lancaster and Souviraà-Labastie, 2020). In line with Choi et al.'s results, extracted vocals are of very satisfying quality – *i.e.*, SDR metric (Vincent et al., 2006) of **9.2dB** for the vocals separated from the HQ-MUSDB test set. In inference, separation is faster than real-time even when running on a CPU on a small laptop.

The isolated vocals are then used to extract the F0 with another neural network, denoted $\mathcal{M}_{\text{F0}}$ in Figure 7.5 and detailed in (Ardaillon and Roebel, 2019).

Musicological expert knowledge

A typical objective is to correlate the voice features with other relevant information and emphasize their relationships via temporal alignment. In the context of singing performances, both lyrics and singing melody (notes) are of high interest. Musicologists have the role to collect such symbolic data and adapt them if necessary to match their research angle.

Regarding the *lyrics* – lyrics are easy to gather for most commercial music today, either online or via an album booklet, so that manual transcription is rarely required. In practice, a succession of *syllables* is more musicologically relevant than *words* for the text as singing notes are held on the vowels of each syllable (Sundberg, 1987). Plus, one can note that the lyrics rarely exactly match the singing content due to additive onomatopoeia (*e.g.*, "yeah", "hm") or unpronounced utterances. Fortunately, it was shown in section 6.4.3 that ADAGIO can handle a *decent* amount of such irregularities and aligns despite missing or additional syllables. The role of the musicologists in lyrics retrieval is thus twofold: (1) to ensure that the text is coherent and correctly written; and (2) to explicitly adapt, whenever necessary, repeated syllables, missing entries, or onomatopoeia judged pertinent, *i.e.*, conveying meaningful interpretative aspects.

Regarding the *melody* – one is looking for the note sequences performed by the singers (as in, *e.g.*, a music score). When the music score is not accessible, transcriptions remain often done by musicologists – this way they can annotate any precise gesture made by the artist.

Voice-to-symbols synchronization

This is the core integration of this thesis into this collaboration since ADAGIO is used to temporally align both syllables and notes.

First, voice-to-syllable alignment is tackled with ADAGIO trained for the acoustic modeling of words – it is denoted $\mathcal{M}_{\text{V2T}}$ in Figure 7.5. Albeit trained for word-level alignment, syllables can be synchronized as well. Indeed, as words and syllables share the same alphabet, the acoustic modeling does not require any adaptation. The only difference lies in the decoding step, as there are more spaces (label ø) to synchronize between syllables than between words.

Second, voice-to-note alignment is tackled with ADAGIO trained for the acoustic modeling of notes – it is denoted $\mathcal{M}_{\text{V2N}}$ in Figure 7.5. A major novel perspective offered by this dual approach is the capability to fully align melisma, *i.e.*, multiple notes held on the same syllable (shown in musical notation by slurs). Musicologists can thus concretely rely on note-level alignment to complement the syllable-level alignment in such case as portrayed in Figure 7.6. To the best of the author' knowledge, it is the first option proposed to musicologists for dealing with the automatic analysis of melisma.

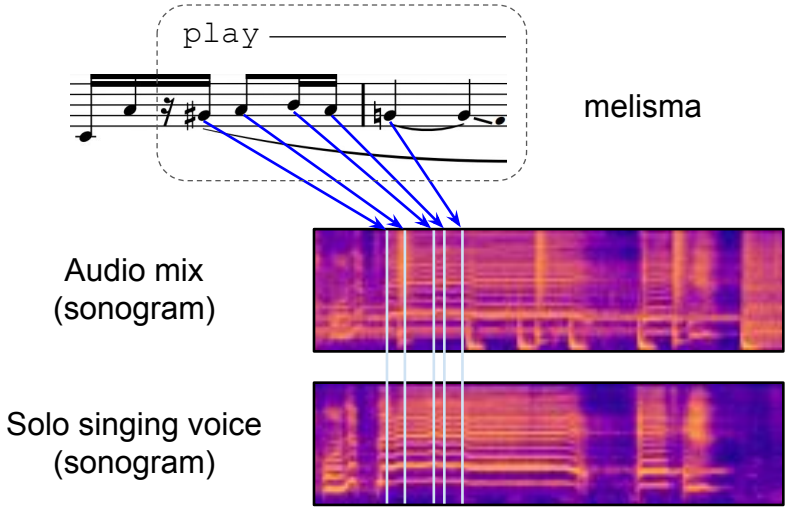


Figure 7.6: Automatic analysis of a melisma: syllable-level alignment only predicts the word "play" on the full duration of this excerpt, without taking pitch variation into account – note-level alignment allows a deeper look into this multi-modal gesture. Score transcription done by Antoine PETIT. See Taylor SWIFT case study – section 7.3.3.

Musicological exploitations

Finally, further musicological studies can be carried out benefiting from the flexibility of the proposed pipeline. The alignment time markers can be read into visualization software (*e.g.*, Sonic Visualiser[3] or RX[4]) for unavoidable manual corrections – being still much less tedious than starting from scratch. All in all, these timestamps are available for qualitative and/or computational analyses. In the rest of this section, a demonstration of the whole musicological protocol is proposed via a case study on Taylor SWIFT. The temporal data from the alignments have allowed performing fine-grained rhythmic analyses and to investigate the structural role played by articulation and micro-rhythm. This is interestingly in line with the previous speech studies (section 7.2) proving interconnections between all voice research angles.

Note that the publication (TEYTAUT et al., 2023) also features a case study on Charles AZNAVOUR, made by co-author Céline CHABOT-CANET, which is focused on vocal phrasing and rhetorical effects involving intonation. However, both for the sake of conciseness and as it mostly exploited the F0 estimation, with textual alignment being only used as a qualitative support to instantaneously relate the effects heard and spectral observations, as shown in Figure 7.7, this analysis is voluntary left out of the current manuscript.

[3] https://www.sonicvisualiser.org
[4] https://www.izotope.com/en/products/rx.html

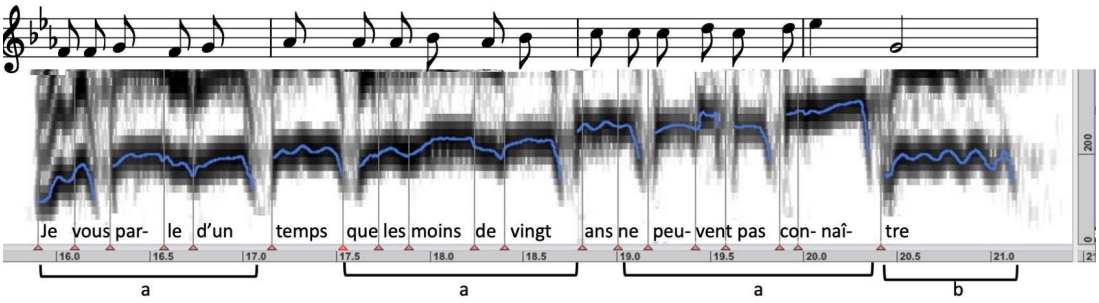


Figure 7.7: Verse 1, line 1 from Charles AZNAVOUR's "La Bohème". Score transcription, sonogram, F0 and text alignment. 'a': fast flow, intonational instability, and absence of vibrato; 'b': sustained note with vibrato on the last syllable of the phrase.

7.3.3 Case study – Taylor SWIFT's "Blank Space"

(This research has been conducted by co-author Antoine PETIT. It is detailed as it demonstrates strong exploitations of alignment time markers and their evaluations via manual corrections.)

This study examines Taylor SWIFT's 2014 hit single "Blank Space"[5] which was selected because (1) at the time of writing, it is SWIFT's second-best charting song and its analysis may uncover what makes a successful song; (2) SWIFT's songs have not yet been the object of much, if any, musicological attention with the recent exception of (PETIT, 2022); and (3) beyond a standard form, its organization reveals lower-level patterns in verses and chorus linked to vocal delivery. Overall, it presents a prime example of analytical work afforded by the pipeline.

The musicologist has focused on the first half of the first verse, *i.e.*, eight bars from `5:30` to `25:30`, a score transcription of which is given in Figure 7.8. It exhibits a **srdc**[6] structure (EVERETT, 2009) and can thus be divided into four parts – two fairly similar segments (**sr**) and a contrasting passage leading to a concluding gesture (**dc**). Following HANNINEN (2012), one may wish to understand which criteria elicit such a quite self-evident segmentation. The claim of this study is that articulation and micro-rhythm play a key role in shaping the form of the excerpt. This can be investigated with the proposed pipeline and its temporal alignments.

Application of the pipeline

Syllables retrieved from a first score transcription were synchronized with the audio with ADAGIO. For each syllable not followed by a rest, the P-center (*perceptual* center, *i.e.*, the moment *heard* as beginning) is approximated by the mean between the start and end times of the syllable. These were then manually checked.

[5] Words and music by Taylor SWIFT, Max MARTIN and SHELLBACK. Reference recording: *1989*, Big Machine, 2014.

[6] **S**tatement, **r**estatement/**r**esponse, **d**eparture, **c**losure.

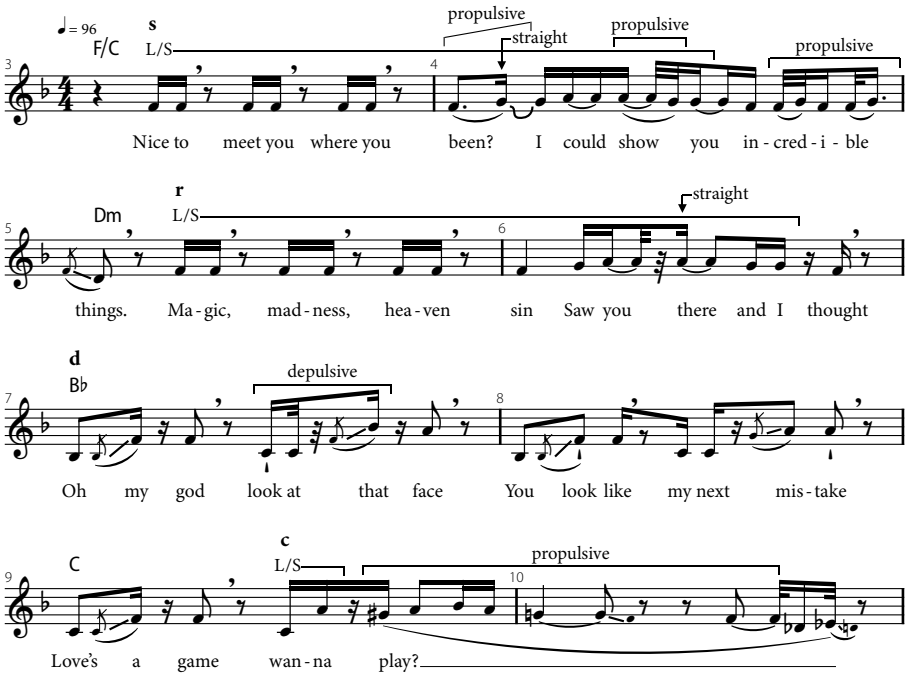


Figure 7.8: Taylor SWIFT, "Blank Space", first verse, measures 3–10. Transcription done by Antoine PETIT.

Next, pitches were transcribed and refined thanks to the visualization of both F0 curves and syllables alignment. The resulting MIDI notes were synchronized to the audio with ADAGIO afterwards. Only the notes associated with the five identified melisma (four in measure 4 and one on "*play?*" in measures 9-10) are kept and their time markers were also manually corrected.

One can then compare the time differences between the corrected markers and the predicted markers leading to an Average Absolute Error (AAE) from a musicological expertise. This is reported on Table 7.1. Interestingly, one can note that the manual syllable corrections are below the theoretical precision $\delta t = 16$ms of ADAGIO in its original setup (which is half of the frame size and given the signal processing setup from Table 4.3). Once again, note alignment appears as less stable but remains acceptable and exploitable.

Alignments	Syllables (*all*)	Notes (on *melisma*)
"Blank Space" (manual correction) ⌞ AAE (ms)	13.5 ± 40.9	22.5 ± 63.0

Table 7.1: Manual alignment corrections for the "Blank Space" case study.

	s	**r**	**d**	**c**
Non-*legato* notes				
⌞ Number	4	6	10	2
⌞ Proportion (%)	20	46	59	22
⌞ Mean duration (ms)	213	162	161	180
std. dev. (ms)	(14.1)	(71.5)	(84.8)	(175)

Table 7.2: Non-*legato* notes by subsection in Taylor SWIFT's "Blank Space", measures 3–10.

Musicological analyses

This timing data at hand − onset, P-center and end of every note in the excerpt − allow analyzing how Taylor SWIFT uses articulation and micro-rhythm to structure her vocals.

Legato articulation, or lack thereof, can be computed by subtracting the onset time of the $n+1^{\text{th}}$ note with the end time of the n^{th} note. The Table 7.2 displays the number, proportion, and mean duration of non-*legato* notes by subsection, painting a vivid picture of form organized through articulation. The excerpt begins with mostly *legato* singing, interspersed with a few very homogeneous silences. SWIFT's vocals then gradually become more jagged − mostly non-*legato*, with many overall shorter, but also much more heterogeneous, silences − before returning to the initial *legato* articulation in the concluding melisma, which is split in two by the longest silence in the excerpt. This arch-like progression seems fairly obvious upon listening (especially when it has been explicitly pointed out beforehand), but may has been missed without the ability to gather accurate timing data.

But articulation is only part of the story. Figure 7.9 maps the duration of every note (*i.e.*, difference between two successive P-centers to which is subtracted the length of the intervening silence, if any) to its P-center. *Legato* articulation is shown with connecting lines, the internal notes of each melisma (aligned with ADAGIO) is shown as unfilled dots. The vertical dashed lines correspond to the beginning of the four subsections, and the horizontal ones to the projected duration of eighth, sixteenth, and thirty-second notes (*i.e.*, the three most frequent symbolic durations in the transcription) at 96BPM.

Not all notes last for their projected duration. In particular, many sixteenth notes appear "uneven", with the on-beat one being longer than the off-beat one. *Long/Short* subdivision (CAPORALETTI, 2014) (*i.e.*, swing) is endemic to **s** and **r**, where it affects *almost* all sixteenth notes, but is absent in **d**, which prioritizes straight eighth notes.

Thus, there exists a subtle interplay between two contrasting *local* vocal styles: (1) mostly *legato*, with L/S (swung) sixteenth notes and step-wise motion (**sr** and **c**); and (2) mostly non-*legato*, with straight rhythms and large leaps filled with gliding intonations (**d**). The more jagged articulation of **r** allows SWIFT to smoothly transition from style (1) to style (2), while the leap of a major sixth coupled with L/S sixteenth notes on "wanna" at the beginning of **c** enables the reverse.

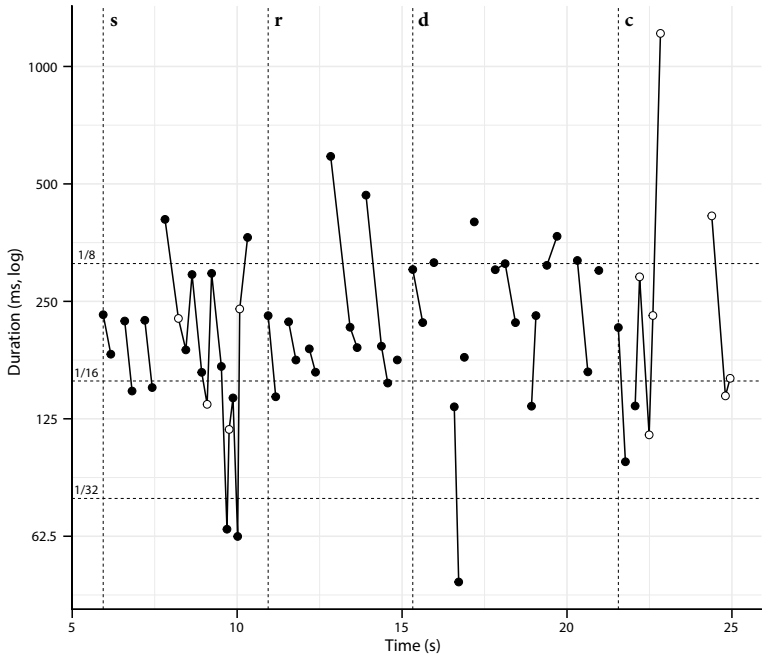


Figure 7.9: Note durations (log-scale) in Taylor Swift's "Blank Space", measures 3–10.

	s	**r**	**d**	**c**
Mean displacement (ms)	32.6	38.6	27.2	30.5
std. dev. (ms)	(38.4)	(27.7)	(9.82)	(29.8)

Table 7.3: Micro-rhythmic displacements by subsection in Taylor Swift's "Blank Space", measures 3–10.

These local styles also share a number of characteristics, among which propulsive tendencies (Caporaletti, 2014) in melismatic passages (*i.e.*, the notes are shorter than projected)[7] and lengthened notes when followed by a silence (both of these can be observed on Figure 7.9), a lack of vibrato (this can be observed on the automatically-extracted F0), and micro-rhythmic displacements of most notes (computed by subtracting their P-center with their projected onset at 96BPM), which consistently appear about 30ms *later* than projected, as shown in Table 7.3.

Thanks to the pipeline streamlining the annotation process, this analysis can easily be expanded upon, so that it encompasses the whole verse/verse-chorus unit/song, etc., up to (at least) the level of the album − thus shedding light on Swift's multifaceted vocal style. The many strategies discovered during the analysis can then be compared with other artists'.

[7]Because such propulsive tendencies are independent from the underlying pulse, which does *not* change, the last note of a melisma must last longer than projected, as "compensation".

Section summary – Musicological pipeline for singing style analysis

A complete pipeline for the musicological analysis of singing voice style has been proposed. It notably exploits deep learning models for singing voice extraction from background music, voice parameter (F0) estimation and – as far as this thesis is concerned – ADAGIO for robust automatic alignment of both syllables *and* notes to the audio. Not only does this pipeline greatly simplify the tedious tasks traditionally done manually by musicologists, but it also offers practical flexibility, as demonstrated in two concrete musicological studies: (1) text and note alignments allowed investigating articulation and micro-rhythm in a Taylor SWIFT song (American pop, 2014); and (2) although not detailed in this manuscript, F0 curves and vowel regions were exploited to highlight vocal phrasing and rhetorical effects involving intonation in a Charles AZNAVOUR song (French chanson, 1966). More generally, this work is meant for future and strong collaborations between musicologists and deep learning researchers sharing a common interest in singing voice. The tools presented will be made available to the community in a web interface hosted at https://passagesxx-xxi.univ-lyon2.fr/activites/projets-anr/projet-ars-analyse-et-transformation-du-style-de-chant-1.

7.4 Perspectives

This thesis paves the way for further investigations, concrete applications and collaborations. This section thus exposes some ongoing projects that cannot be considered as achieved at the time of writing but that are in line with the main scope of this thesis, as well as burgeoning ideas for future research.

Automatic corpora segmentation

The alignment of very long audio recordings has been a well-known limitation to many aligners (BORDEL et al., 2015; KATSAMANIS et al., 2011). This is due to classical DTW- or CTC-based decoding modules $\mathcal{D}_\pi$, which compute the forced alignment *per se*, as they require the computation and storage of an alignment path that scales *quadratically* in number of frames and symbols to synchronize (cf. section 2.3.1). This quickly becomes prohibitive for long audio and symbolic sequences – exceeding the memory storage available on modern machines.

This thesis was involved in a project on this issue (DORAS et al., 2023) in which ADAGIO was the acoustic model $\mathcal{M}_\Theta$ used in experiments and was revealed for the first time. However, as the main contribution of the paper was the design a new decoding module $\mathcal{D}_\pi$ supporting very long sequences, and was implemented by the paper's first author, this section is voluntarily kept short. The core idea was to adapt the CTC decoding with an exact reformulation of the DTW algorithm, which scales linearly in memory, introduced by TRALIE and DEMPSEY (2020).

This aligner was shown to be capable of synchronizing several hours of audio and text even for different languages and despite some errors in transcripts (as the evaluations presented in section 6.4.3). In the future, it is planned to use this system for the automatic segmentation of large corpora like audiobooks into small excerpts of paired audio and labels for enriching and completing existing datasets for downstream tasks, such as voice recognition, conversion or synthesis (Kürzinger et al., 2020; Roebel and Bous, 2022).

Joint singing voice alignment and separation

An open research problem is to investigate to what extent both separation and alignment can be addressed simultaneously. These tasks are traditionally studied independently, but recent research shed light on their potential complementarity.

Indeed, Stoller et al. used the same neural architecture to perform each task individually (Stoller et al., 2018, 2019); Schulze-Forster et al. (2021) derived alignment from an attention mechanism informing voice separation; Meseguer-Brocal and Peeters (2020) exploited pre-aligned phonemes to help vocal extraction; Vaglio et al. (2020b) built their acoustic model on pre-isolated vocals; and this thesis improved voice alignment thanks to an auxiliary constraint of spectral reconstruction of estimated vocals.

These results tend to indicate that alignment and separation are beneficial to each other. Yet, to the best of the author's knowledge, no approach has tried a joint training procedure so far. One potentially interesting direction to start with would be the work of Choi et al. (2019) that proposed a relatively light U-Net architecture for singing voice extraction in the spectral domain. (It was the voice separator used for the temporal constraints in section 5.2 and the musicological pipeline from section 7.3.) The continuation of this work (Choi et al., 2021) introduced a signal-based conditioning that might be adapted for alignment data.

Note alignment for voice and music

Further experiments are required to enhance and complete the study on the synchronization between notes and singing voice as already mentioned in section 6.4.2. The author believes there is way for improvement in terms of alignment accuracy and, also, in exploitation of the posteriorgram for singing note transcription based on *empirical* observations of the saliency maps – see Figure 6.4. In a similar vein, the problem of note alignment can go beyond voice signals such that the alignment of monophonic instruments could probably be handled with a CTC-based aligner like ADAGIO. To go even further, one could tackle the simultaneous alignment of multiple notes for polyphonic instruments by taking inspiration, *e.g.*, from the work of Weiss and Peeters (2021) who introduced a multi-label CTC for polyphonic *transcription*. A final objective would be to integrate such CTC-based acoustic modeling into a real-time audio-to-score alignment (Lajugie et al., 2016) for score following applications (Cont, 2011).

Improving the acoustic modeling & transcription

This work, as commonly done in the alignment literature, has extensively relied on acoustic modeling approaches originally thought for voice transcription to decode and synchronize symbols to audio from some hidden representations. As a result, it is reasonable to assume that better acoustic modeling will lead to performance improvements. Two options in that direction come to mind. Either to keep a flexible and general end-to-end architecture and informing it with other types of information *e.g.*, pitch detection as recently proposed by Huang et al. (2022), or to add additional modules, like language models (Ycart et al., 2019), to complement the acoustic model in the decoding step – this is the strategy supported by Gupta et al. (2020). It is worth mentioning that this thesis was concerned with the former strategy, *i.e.*, additional temporal constraints. The later strategy would contradict the thesis initial objective, which was to be able to align without domain knowledge. More generally, there is an aim for universality such that a theoretically optimal system could proceed to both audio/voice transcription and alignment – even without requiring the ground-truth texts (Zhu et al., 2022).

Section summary – Perspectives

There are many upcoming investigations in the continuation of this work as voice alignment algorithms can (1) serve to automatically segment large corpora; (2) be coupled with the voice separation task; (3) be improved and extended for notes and melody synchronization; and (4) keep motivating the search of an "universal" transcription-alignment system.

7.5 Applications and collaborations in a nutshell

Chapter summary – Applications and collaborations

In this chapter, the practical applications of my thesis were presented by means of conducted collaborations. First, I investigated singing voice synthesis in a concatenative strategy for which ADAGIO was trained to voluntary specialize (overfit) on a new recorded dataset and produce phonetic alignments. Then, I detailed collaborative studies on voice expressivity focused on production strategies of both spoken social attitudes and singing style in musical performances. The former (speech) revealed for the first time that social attitudes are also conveyed through articulation. The later (singing) introduced a complete pipeline dedicated to the analysis of singing style with neural voice processing and alignment of both syllables and notes, that has already been exploited by musicology partners. Lastly, I mentioned recent ongoing works and perspectives for future research – notably one following a project for the alignment of very long audio recordings.

Chapter 8

Conclusion

"Never is an awfully long ***Time****."*

Peter Pan
– James Matthew Barrie

Time.

An essential, and yet so perplexing, aspect of the universe and human daily life. Many physical laws are governed by time so that it determines the evolution of various observations and surrounding signals. Audio data are no exception: music and voice are fundamentally non-stationary signals whose temporal study can reveal many properties.

Humans can interact with one another through messages that they can commonly express with their voices, as *orality* is a primary modality of human communication. These vocal locutions are very often associated with underlying representations of *symbolic* nature, which translate another modality of communication, typically depending on a textual language (*e.g.*, stories, tales, poems, legends, etc.) and/or music theory (*e.g.*, songs, opera, musical scores).

In the context of such communicative approaches, the notion of temporality is of utmost importance and has greatly motivated research studies dedicated to its analysis. This thesis has precisely proposed to automatically tackle the problem of temporal alignment between voice signals and their related symbolic sequences, to uncover some key temporal aspects involved in human communication. In doing so, it was aimed to bridge, to some extent, the *semantic gap* between diverse communication modalities.

8.1 Manuscript summary

The content of this thesis is briefly summarized in this section.

(Chapter 2.) First and foremost, I exposed the overall context of my work: to better understand some expressive strategies in human communication by revealing the temporal relationships between two voice modalities − oral and symbolic. To characterize the orality, I introduced signal processing basics allowing to highlight relevant voice features capturing the *temporal* evolution of the spectral content. The representation of symbolic sequences was highly correlated with the history of writing of both spoken languages and music theory. At the heart of the interaction between voice signals and symbols, lie two prominent analysis tasks: transcription and alignment. My thesis has been primarily interested in temporal voice alignment, or synchronization, which aims at associating each symbol with a timestamp of appearance in the audio recording, and offers thrilling applications for the general public (automatic closed-captioning, karaoke) and research community (voice analysis and synthesis).

(Chapter 3.) Through a systemic review of the literature, I showed that any system for temporal voice alignment is composed of (1) an *acoustic model*, inferring a time-symbol representation from the audio, and (2) a *decoding module*, using this representation for the forced alignment of a ground-truth sequence. While the decoding module is typically based on a reference dynamic programming algorithm, the acoustic model has seen many implementations. Inspired by recent advances in deep learning for voice processing, I motivated the choice to design the acoustic model as a deep neural network. In particular, I presented the Connectionist Temporal Classification (CTC) framework, which is particularly convenient and appealing for end-to-end voice alignment and was the chosen approach for my thesis.

(Chapter 4.) Then, I presented my preliminary proposals of CTC-based aligners which allowed me to develop a series of criteria that a robust and practical aligner should respect. These criteria were not systematically validated either by existing systems or by my first attempts, which prompted the development of a new acoustic model in accordance with these identified needs. This is ADAGIO − for **A**utomatic **D**eep **A**li**G**nment of v**OI**ce −, the central node of my thesis research. Trained on the CTC criterion, ADAGIO predicts, from an audio, a posteriorgram estimating the temporal evolution of the underlying symbolic content.

(Chapter 5.) However, since the CTC is originally a transcription measure, it can be minimized without aligning as precisely as desired. Faced with this intrinsic limitation, I proposed to introduce additional temporal constraints when training the networks to ensure the emergence of alignment properties directly in the posteriorgrams. This temporal knowledge took the form of ancillary tasks of spectral reconstruction, temporal structure propagation, and time-sequence monotonicity. This defined a version of ADAGIO enhanced with temporal information learning. It is worth highlighting that these enhancements were purposely thought generic, *i.e.* not based on any domain/expert knowledge (*e.g.*, specific language or music genre).

(Chapter 6.) Objective evaluations based on classical metrics were conducted to quantify, and ultimately confirm, the suitability of ADAGIO for voice-to-text and (to a lesser extent) voice-to-note alignments, and the beneficial impacts of my auxiliary temporal constraints when supervised on isolated voices. Two word-level aligned datasets were derived from existing works: *Philos 10* for speech and *Playlist 50* for lyrics and notes. Although state of the art is not reached in the classical evaluation case, ADAGIO stands out for its simplicity and diversity of use.

(Chapter 7.) Finally, ADAGIO has been put into practice through various collaborations in which I performed the alignments and analyses were dominantly done by respective end users. These works have focused on the concatenative synthesis of singing voice via phonetic alignment, and the study of expressive production strategies for both spoken social attitudes and singing style in musical performances − bridging, to some extent, the *semantic gap* between symbols and audio. Lastly, ongoing research and perspectives were mentioned.

8.2 Coming full circle: back to initial questions

In the course of this thesis, I have *deeply learned* to investigate and answer to practical and research questions − and notably the ones raised by temporal voice alignment in section 1.1. To give a greater overview to this research, and essentially conclude this manuscript, this last section sums up the answers I propose to these initial questions from accumulated knowledge.

Question 1 (Q1). *Temporal voice alignment – what?*

- *What* is the temporal alignment task in general and for voice?

  A task consisting in predicting a precise timestamping for each event reported in a sequence of ordered elements expressed in an audio recording. This thesis specializes on voice alignment, which links oral and symbolic modalities of human communication.

- *What* kind of representations can be used to align voice data?

  Time-Frequency Representations (TFR) are relevant features to manipulate voice signals as they capture the variations of the spectral content over time. Symbolic sequences are defined through an associated finite alphabet of labels (*e.g.*, letters, digits, notes, etc.).

Question 2 (Q2). *Temporal voice alignment – how?*

- *How* to develop a system for the temporal alignment of voice?

  Most voice aligners in the literature are composed of an acoustic model and a decoding module. The decoding module often relies on a Dynamic Time Warping (DTW) algorithm. Recent acoustic models exploit deep neural networks learning robust features from audio. In this thesis, an acoustic model, ADAGIO, was proposed as an end-to-end, convolutional network trained in the highly flexible Connectionist Temporal Classifical (CTC) framework.

- *How* to extract suitable temporal information from audio to reinforce alignment accuracy?

  The main claim of this research is that incorporating additional temporal information in the acoustic modeling must result in higher alignment accuracy. This appeared as a natural extension for ADAGIO, as the flexibility offered by the CTC came with no intrinsic guarantee of precise alignments. In the end, temporal spectral reconstruction, temporal structure propagation, and audio-symbolic monotonicity were introduced as new supplementary and general (*i.e.* without domain knowledge) objectives that were integrated in the training phase. These constraints, when supervised on solo vocals estimated with voice separation, were shown to have a beneficial impact on voice alignment.

Question 3 (Q3). *Temporal voice alignment – why?*

- *Why* is temporal alignment of interest in various research communities?

  General alignment algorithms emerged in many fields requiring similarity measures between sequences such as telecommunications, bioinformatics and the audio domain. The specific case of voice synchronization appears as a means to bridge, to some extent, the semantic gap between different modalities involved in human communication by uncovering and looking closely at the temporal relationships between these diverse representations. Voice aligners have already found well-known mainstream applications for the general public such as text-based audio indexing, automatic closed captioning, or karaoke generation.

- *Why* does temporal voice alignment lead to numerous research applications?

  Beyond the above-mentioned use cases, temporal voice alignment is of great interest in research dedicated to voice analysis and synthesis. Concatenative singing synthesis, indeed, exploits aligned phonetic content to generate sung utterances. The automatic segmentation of audiobooks, often needed by speech specialists, is possible via long audio alignment. Synchronizing voice data also allows fine-grained analysis of voice expressivity and thus a better understanding of, *e.g.*, the production strategies at play to convey social attitudes or singing interpretative style according to musicological studies.

Naturally, one could wonder whether asking such questions in the first place was of interest. In the very words of Luciano Berio (2006), "*I think that the search for a universal answer to the questions raised by musical experience will never be completely fulfilled; but we know that a question raised is often more significant than the answer received. Only a reckless spirit, today, would try to give a total explanation of music, but anyone who would never pose the problem is even more reckless.*" Perhaps the very essence of questioning music and voice, as essential expressive and communicative vectors, would be a springboard to better apprehend *time* itself.

And even beyond music, voice and time − the very act of raising questions, of all kinds and at all levels, seems more essential than ever, today, to stimulate the minds and the intellect. So many more questions have to be raised to, hopefully, face the challenges of the century such as global warming, modern wars and human condition − this sure will require some *time*.

End of the Thesis

"***Time*** isn't the main thing. It's the only thing."

Miles Davis

List of Figures

List of Tables

Bibliography

J. Aach and G. M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 2001. (Cited on page 50.)

J. Abesser and M. Müller. Jazz bass transcription using a u-net architecture. *Electronics*, 10(6):670, 2021. (Cited on page 48.)

L. Ardaillon. *Synthesis and expressive transformation of singing voice (Chapter 3)*. PhD thesis, Paris VI, 2017. (Cited on pages 42, 112, 130, and 154.)

L. Ardaillon and A. Roebel. Fully-convolutional network for pitch estimation of speech signals. In *Insterspeech 2019*, 2019. (Cited on pages 42 and 138.)

L. Ardaillon, C. Chabot-Canet, and A. Roebel. Expressive control of singing voice synthesis using musical contexts and a parametric f0 model. In *Interspeech 2016*, volume 2016, pages 1250–1254, 2016. (Cited on pages 32, 130, 132, and 135.)

P. Arias, P. Belin, and J.-J. Aucouturier. Auditory smiles trigger unconscious facial imitation. *Current Biology*, 28(14):R782–R783, 2018a. (Cited on page 32.)

P. Arias, C. Soladie, O. Bouafif, A. Roebel, R. Seguier, and J.-J. Aucouturier. Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*, 11(3):507–518, 2018b. (Cited on page 32.)

A. Arzt. *Flexible and robust music tracking*. PhD thesis, Johannes Kepler University, Linz, 2016. (Cited on page 67.)

A. Arzt, G. Widmer, and S. Dixon. Automatic page turning for musicians via real-time machine listening. In *ECAI 2008*, pages 241–245. IOS Press, 2008. (Cited on page 50.)

D. A. Backstrom, B. V. Tucker, and M. C. Kelley. Forced-alignment of the sung acoustic signal using deep neural nets. *Canadian Acoustics*, 47(3):98–99, 2019. (Cited on page 70.)

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. (Cited on page 61.)

C. Banderier and S. Schwer. Why delannoy numbers? *Journal of statistical planning and inference*, 135(1):40–54, 2005. (Cited on page 65.)

K. Bantupalli and Y. Xie. American sign language recognition using deep learning and computer vision. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4896–4899. IEEE, 2018. (Cited on page 47.)

H. B. Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989. (Cited on page 58.)

A. Baró, P. Riba, J. Calvo-Zaragoza, and A. Fornés. From optical music recognition to handwritten music recognition: a baseline. *Pattern Recognition Letters*, 123:1–8, 2019. (Cited on page 48.)

Y. Bengio, P. Frasconi, and P. Simard. The problem of learning long-term dependencies in recurrent networks. In *IEEE international conference on neural networks*, pages 1183–1188. IEEE, 1993. (Cited on page 61.)

L. Berio. *Remembering the future*, volume 52. Harvard University Press, 2006. (Cited on page 150.)

M. Bernard and H. Titeux. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958, 2021. (Cited on page 133.)

M. Besson and D. Schön. Comparison between language and music. *Annals of the New York Academy of Sciences*, 930(1):232–258, 2001. (Cited on page 26.)

R. M. Bittner, J. Salamon, S. Essid, and J. P. Bello. Melody extraction by contour classification. In *International Conference on Music Information Retrieval (ISMIR)*, 2015. (Cited on page 48.)

T. Bluche, H. Ney, J. Louradour, and C. Kermorvant. Framewise and ctc training of neural networks for handwriting recognition. *13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 81–85, 2015. (Cited on page 91.)

G. Bordel, M. Penagarikano, L. J. Rodríguez-Fuentes, A. Álvarez, and A. Varona. Probabilistic kernels for improved text-to-speech alignment in long audio tracks. *IEEE Signal Processing Letters*, 23(1):126–129, 2015. (Cited on page 144.)

J.-P. Briot. Apprentissage profond et génération de musique. *ActuIA–Le magazine de l'intelligence artificielle,(7)*, 46:2682–1265, 2022. (Cited on page 55.)

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer, et al. The mathematics of statistical machine translation: Parameter estimation. 1993. (Cited on page 50.)

F. Brugnara, D. Falavigna, and M. Omologo. Automatic segmentation and labeling of speech based on hidden markov models. *Speech Communication*, 12(4):357–370, 1993. (Cited on page 67.)

D. A. Brunow and T. A. Cullen. Effect of text-to-speech and human reader on listening comprehension for students with learning disabilities. *Computers in the Schools*, 38(3):214–231, 2021. (Cited on page 42.)

J. Calvo-Zaragoza, J. J. Valero-Mas, and A. Pertusa. End-to-end optical music recognition using neural networks. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR*, pages 23–27, 2017. (Cited on page 48.)

V. Caporaletti. *Swing e Groove: sui fondamenti estetici delle musiche audiotattili*. Libreria Italiana Musicale, 2014. (Cited on pages 142 and 143.)

C. Chabot-Canet. *Léo Ferré: une voix et un phrasé emblématiques*. L'Harmattan, 2008. (Cited on page 136.)

C. Chabot-Canet. L'analyse spectrale au fondement d'une rhétorique des styles interprétatifs dans la chanson française. *Volume!*, 16(2)/17(1):29–47, 2020a. (Cited on page 136.)

C. Chabot-Canet. Spectrum analysis and the rhetorics of performance styles in french chanson. *Volume!*, 162171(1):29–47, 2020b. (Cited on page 32.)

C. Chabot-Canet, L. Ardaillon, and A. Roebel. Modeling vocal styles, synthesizing expressive singing. the example of édith piaf. *Volume!*, 162171(1):63–85, 2020. (Cited on pages 32 and 40.)

Y.-P. Cho, F.-R. Yang, Y.-C. Chang, C.-T. Cheng, X.-H. Wang, and Y.-W. Liu. A survey on recent deep learning-driven singing voice synthesis systems. In *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 319–323. IEEE, 2021. (Cited on page 41.)

W. Choi, M. Kim, J. Chung, D. Lee, and S. Jung. Investigating U-Nets with various intermediate blocks for spectrogram-based singing voice separation. *arXiv preprint arXiv:1912.02591*, 2019. (Cited on pages 93, 117, 137, and 145.)

W. Choi, M. Kim, J. Chung, and S. Jung. LaSAFT: Latent source attentive frequency transformation for conditioned source separation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175. IEEE, 2021. (Cited on page 145.)

C. S. Chong, J. Kim, and C. Davis. Disgust expressive speech: the acoustic consequences of the facial expression of emotion. *Speech Communication*, 98:68–72, 2018. (Cited on page 32.)

A. Cohen-Hadria and G. Peeters. Music structure boundaries estimation using multiple self-similarity matrices as input depth of convolutional neural networks. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017. (Cited on page 94.)

A. Cohen-Hadria, A. Roebel, and G. Peeters. Improving singing voice separation using deep u-net and wave-u-net with data augmentation. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019. (Cited on pages 42, 93, and 137.)

R. Collier and J. Hart. The role of intonation in speech perception. In *Structure and process in speech perception*, pages 107–123. Springer, 1975. (Cited on page 32.)

R. Collobert, C. Puhrsch, and G. Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016. (Cited on pages 71 and 84.)

A. Cont. A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):974–987, 2009. (Cited on pages 50 and 68.)

A. Cont. On the creative use of score following and its impact on research. In *SMC 2011: 8th Sound and Music Computing conference*, 2011. (Cited on page 145.)

A. Cont, D. Schwarz, N. Schnell, and C. Raphael. Evaluation of real-time audio-to-score alignment. In *International Society for Music Information Retrieval (ISMIR)*, 2007. (Cited on pages 112 and 113.)

A. Cont, J. Echeveste, J.-L. Giavitto, and F. Jacquemard. Correct automatic accompaniment despite machine listening or human errors in antescofo. In *ICMC 2012-International Computer Music Conference*, 2012. (Cited on page 50.)

G. Cook. *Applied linguistics*. Oxford University Press, 2003. (Cited on page 26.)

E. E. Cust, A. J. Sweeting, K. Ball, and S. Robertson. Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. *Journal of sports sciences*, 37(5):568–600, 2019. (Cited on page 47.)

P. Cuvillier. *On temporal coherency of probabilistic models for audio-to-score alignment*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2016. (Cited on page 68.)

P. Cuvillier and A. Cont. Coherent time modeling of semi-markov models with application to real-time audio-to-score alignment. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2014. (Cited on page 68.)

R. B. Dannenberg. An on-line algorithm for real-time accompaniment. In *ICMC*, volume 84, pages 193–198, 1984. (Cited on page 67.)

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980. (Cited on page 38.)

E. Demirel, S. Ahlbäck, and S. Dixon. Mstre-net: Multistreaming acoustic modeling for automatic lyrics transcription. *International Society for Music Information Retrieval (ISMIR)*, 2021. (Cited on page 111.)

A. Dister, M. Constant, and G. Purnelle. Normalizing speech transcriptions for natural language processing. In *3rd International Conference on Spoken Communication (GSCP'09)*, 2009. (Cited on page 47.)

S. Dixon. An on-line time warping algorithm for tracking musical performances. In *IJCAI*, pages 1727–1728, 2005. (Cited on page 67.)

G. Doras, Y. Teytaut, and A. Roebel. A linear memory ctc-based algorithm for text-to-voice alignment of very long audio recordings. *Applied Sciences*, 13(3):1854, 2023. (Cited on pages 22, 110, 111, 112, 124, 136, and 144.)

C. Douwes, G. Bindi, A. Caillon, P. Esling, and J.-P. Briot. Is quality enough? integrating energy consumption in a large-scale evaluation of neural audio synthesis models. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023. (Cited on page 55.)

Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang. The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9. IEEE, 2013. (Cited on pages 40, 41, 68, and 152.)

H. Dudley, R. R. Riesz, and S. S. Watkins. A synthetic speaker. *Journal of the Franklin Institute*, 227(6):739–764, 1939. (Cited on page 41.)

D. P. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1): 51–60, 2007. (Cited on page 67.)

W. Everett. *The Foundations of Rock: From "Blue Suede Shoes" to "Suite: Judy Blue Eyes"*. Oxford University Press, 2009. (Cited on page 140.)

M. Fares, C. Pelachaud, and N. Obin. Multimodal generation of upper-facial and head gestures with a transformer network using speech and text. 2021. (Cited on page 27.)

S. Farner, A. Roebel, C. Veaux, G. Beller, X. Rodet, and L. Ach. Voice transformation and speech synthesis for video games. In *Paris Game Developers Conference, Paris, France*, 2008. (Cited on page 42.)

H. Fastl and E. Zwicker. *Psychoacoustics: facts and models*, volume 22. Springer Science & Business Media, 2006. (Cited on page 37.)

M. Fell, Y. Nechaev, G. Meseguer-Brocal, E. Cabrio, F. Gandon, and G. Peeters. Lyrics segmentation via bimodal text–audio representation. *Natural Language Engineering*, 28(3):317–336, 2022. (Cited on page 94.)

S. R. Fischer. *History of writing*. Reaktion books, 2003. (Cited on page 27.)

D. Folgado, M. Barandas, R. Matias, R. Martins, M. Carvalho, and H. Gamboa. Time alignment measurement for time series. *Pattern Recognition*, 81:268–279, 2018. (Cited on page 50.)

I. Fónagy. *La Vive Voix: essais de psycho-phonétique*. Payot, 1983. (Cited on page 136.)

M. Freitag and Y. Al-Onaizan. Beam search strategies for neural machine translation. *ACL 2017*, page 56, 2017. (Cited on page 45.)

H. Fujihara and M. Goto. Lyrics-to-audio alignment and its application. In *Dagstuhl Follow-Ups*, volume 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012. (Cited on page 51.)

X. Gao, C. Gupta, and H. Li. Lyrics transcription and lyrics-to-audio alignment with music-informed acoustic models. *MIREX*, 2021. (Cited on pages 69 and 115.)

X. Gao, C. Gupta, and H. Li. Automatic lyrics transcription of polyphonic music with lyrics-chord multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2280–2294, 2022. (Cited on page 136.)

D. Garreau, R. Lajugie, S. Arlot, and F. Bach. Metric learning for temporal sequence alignment. *Advances in neural information processing systems*, 27, 2014. (Cited on page 65.)

U. Germann. Greedy decoding for statistical machine translation in almost linear time. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 72–79, 2003. (Cited on page 45.)

R. Gong, P. Cuvillier, N. Obin, and A. Cont. Real-time audio-to-score alignment of singing voice based on melody and lyric information. In *Interspeech*, 2015. (Cited on page 68.)

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016. (Cited on page 55.)

I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *International Conference on Learning Representations*, 2014. (Cited on page 47.)

K. Gorman, J. Howell, and M. Wagner. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193, 2011. (Cited on page 68.)

D. Gottlieb and C.-W. Shu. On the gibbs phenomenon and its resolution. *SIAM review*, 39 (4):644–668, 1997. (Cited on page 36.)

A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR, 2014. (Cited on page 71.)

A. Graves, S. Fernández, F. Gomez, and J. 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. (Cited on pages 71, 72, 74, 75, and 90.)

A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013. (Cited on page 69.)

L. Grubb and R. B. Dannenberg. Automated accompaniment of musical ensembles. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI)*, pages 94–99, 1994. (Cited on page 67.)

C. Gupta, D. Grunberg, P. Rao, and Y. Wang. Towards automatic mispronunciation detection in singing. In *ISMIR*, pages 390–396, 2017. (Cited on page 67.)

C. Gupta, H. Li, and Y. Wang. Automatic pronunciation evaluation of singing. In *Interspeech*, pages 1507–1511, 2018. (Cited on page 115.)

C. Gupta, E. Yilmaz, and H. Li. Automatic lyrics alignment and transcription in polyphonic music: Does background music help? In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 496–500. IEEE, 2020. (Cited on pages 69, 115, and 146.)

D. A. Hanninen. *A Theory of Music Analysis: On Segmentation and Associative Organization*. University of Rochester Press, 2012. (Cited on page 140.)

A. Hannun. Sequence modeling with ctc. *Distill*, 2017. doi: 10.23915/distill.00008. https://distill.pub/2017/ctc. (Cited on pages 75, 90, 101, and 153.)

A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, Sengupta Shubho, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. (Cited on page 69.)

J. K. Hansen. Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients. In *9th Sound and Music Computing Conference (SMC)*, pages 494–499, 2012. (Cited on pages 39, 111, and 152.)

A. M. Hein. Identification and bridging of semantic gaps in the context of multi-domain engineering. In *Forum on Philosophy, Engineering & Technology*, pages 58–57, 2010. (Cited on page 16.)

C. F. Hellwag. *Dissertatio de Formatione Loquelae (1781)*. Verlag von Gebr. Henninger, 1886. (Cited on page 31.)

R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5 (50):2154, 2020. (Cited on page 85.)

N. Henrich Bernardoni. La voix timbrée dans les chansons: considérations physiologiques et acoustiques. *Volume!*, 16(2)/17(1):49–61, 2020. (Cited on page 136.)

J. R. Higgins et al. *Sampling theory in Fourier and signal analysis: foundations*. Oxford University Press on Demand, 1996. (Cited on page 34.)

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012. (Cited on page 63.)

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997. (Cited on page 61.)

T. Hori, S. Watanabe, Y. Zhang, and W. Chan. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. *Proc. Interspeech 2017*, pages 949–953, 2017. (Cited on page 71.)

J.-P. Hosom. Speaker-independent phoneme alignment using transition-dependent states. *Speech communication*, 51(4):352–368, 2009. (Cited on page 68.)

J. Huang, E. Benetos, and S. Ewert. Improving lyrics alignment through joint pitch detection. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 451–455. IEEE, 2022. (Cited on page 146.)

E. J. Humphrey, J. P. Bello, and Y. LeCun. Moving beyond feature design: Deep architectures and automatic feature learning in music informatics. In *International Society for Music Information Retrieval (ISMIR)*, pages 403–408, 2012. (Cited on page 55.)

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. (Cited on page 62.)

F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on acoustics, speech, and signal processing*, 23(1):67–72, 1975. (Cited on pages 44 and 66.)

C. Jacques and A. Roebel. Data augmentation for drum transcription with convolutional neural networks. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019. (Cited on page 48.)

A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde. Singing voice separation with deep u-net convolutional networks. 2017. (Cited on page 93.)

J. Jiang, K. Chen, W. Li, and G. Xia. Large-vocabulary chord transcription via chord structure decomposition. In *International Society for Music Information Retrieval (ISMIR)*, pages 644–651, 2019. (Cited on page 48.)

P. N. Juslin. *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA, 2019. (Cited on page 32.)

A. Katsamanis, M. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan. Sailalign: Robust long speech-text alignment. In *Proceedings of workshop on new tools and methods for very-large scale phonetics research*, 2011. (Cited on page 144.)

M. C. Kelley and B. V. Tucker. A comparison of input types to a deep neural network-based forced aligner. 2018. (Cited on page 70.)

R. D. Kent and C. Read. *The acoustic analysis of speech*. San Diego, California: Singular, 2nd edn, 2002. (Cited on page 40.)

S. Kim, T. Hori, and S. Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE, 2017. (Cited on pages 71, 90, and 101.)

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 59.)

H. Kirchhoff and A. Lerch. Evaluation of features for audio-to-audio alignment. *Journal of New Music Research*, 40(1):27–41, 2011. (Cited on page 50.)

L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings 22*, pages 267–278. Springer, 2020. (Cited on pages 71 and 145.)

S. Lacasse. The phonographic voice: paralinguistic features and phonographic staging in popular music singing. In A. Bayley, editor, *Recorded Music: Performance, Culture and Technology*, pages 225–251. Cambridge University Press, 2010. (Cited on page 136.)

A. Lacheret-Dujour and F. Beaugendre. *La Prosodie du français*. CNRS, 1999. (Cited on page 136.)

R. Lajugie, P. Bojanowski, P. Cuvillier, S. Arlot, and F. Bach. A weakly-supervised discriminative model for audio-to-score alignment. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*, pages 2484–2488. IEEE, 2016. (Cited on page 145.)

E. P. Lancaster and N. Souviraà-Labastie. A frugal approach to music source separation, 2020. URL https://hal.science/hal-02986241v1/document. (Cited on pages 93 and 137.)

P. Lanchantin, A. C. Morris, X. Rodet, and C. Veaux. Automatic phoneme segmentation with relaxed textual constraints. In *LREC*, 2008. (Cited on page 131.)

C. Le Moine and N. Obin. Att-hack: An expressive speech database with social attitudes. *Speech Prosody,* Tokyo, Japan, May, 2020. (Cited on pages 32, 110, and 133.)

C. Le Moine, N. Obin, and A. Roebel. Speaker attentive speech emotion recognition. In *Interspeech 2021*, pages 2866–2870. ISCA, 2021. (Cited on page 32.)

Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989. (Cited on pages 47, 48, and 153.)

Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. (Cited on page 60.)

H. Leung and V. Zue. A procedure for automatic alignment of phonetic transcriptions with continuous speech. In *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 73–76. IEEE, 1984. (Cited on page 67.)

K.-W. Liang, H.-C. Lee, Y.-T. Lai, and P.-C. Chang. Query by singing and humming system based on combined dtw and linear scaling. In *2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2. IEEE, 2021a. (Cited on page 42.)

S. Liang, C. Deng, and Y. Zhang. A simple approach to balance task loss in multi-task learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 812–823. IEEE, 2021b. (Cited on page 101.)

A. H. Liu, W.-N. Hsu, M. Auli, and A. Baevski. Towards end-to-end unsupervised speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 221–228. IEEE, 2023. (Cited on page 48.)

H. Liu, S. Jin, and C. Zhang. Connectionist temporal classification with maximum entropy regularization. *Advances in Neural Information Processing Systems*, 31:831–841, 2018. (Cited on page 91.)

A. Liutkus, D. Fitzgerald, and Z. Rafii. Scalable audio separation with light kernel additive modelling. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 76–80. IEEE, 2015. (Cited on page 137.)

S. R. Livingstone, K. Peck, and F. A. Russo. Acoustic differences in the speaking and singing voice. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 035080. Acoustical Society of America, 2013. (Cited on page 40.)

A. Ljolje and M. Riley. Automatic segmentation and labeling of speech. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pages 473–476. IEEE Computer Society, 1991. (Cited on pages 67 and 68.)

R. Loiseau, T. Monnier, M. Aubry, and L. Landrieu. Representing shape collections with alignment-aware linear models. In *2021 International Conference on 3D Vision (3DV)*, pages 1044–1053. IEEE, 2021. (Cited on page 55.)

R. Loiseau, B. Bouvier, Y. Teytaut, E. Vincent, M. Aubry, and L. Landrieu. A model you can hear: Audio identification with playable prototypes. In *23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 2022. (Cited on page 22.)

I. López-Espejo, Z.-H. Tan, J. H. Hansen, and J. Jensen. Deep spoken keyword spotting: An overview. *IEEE Access*, 10:4169–4199, 2021. (Cited on page 42.)

P. Léon. *Précis de phonostylistique: parole et expressivité*. Nathan Université, 1993. (Cited on page 136.)

A. L. Maas, A. Y. Hannun, A. Y. Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 30, page 3. Atlanta, Georgia, USA, 2013. (Cited on page 56.)

I. P. Machado, A. L. Gomes, H. Gamboa, V. Paixão, and R. M. Costa. Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization. *Information Processing & Management*, 51(2):204–214, 2015. (Cited on page 50.)

L. Mao. Number of alignments in connectionist temporal classification (ctc). web resource, https://leimao.github.io/blog/CTC-Alignment-Combinations/, 2019. accessed 18 March 2023. (Cited on page 101.)

N. L. Masclef, A. Vaglio, and M. Moussallam. User-centered evaluation of lyrics-to-audio alignment. In *International Society for Music Information Retrieval (ISMIR)*, pages 420–427, 2021. (Cited on page 113.)

P. McAleer, A. Todorov, and P. Belin. How do you say 'hello'? personality impressions from brief novel voices. *PloS one*, 9(3):e90779, 2014. (Cited on page 133.)

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017. (Cited on page 68.)

B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015. (Cited on page 85.)

G. Meseguer-Brocal and G. Peeters. Content based singing voice source separation via strong conditioning using aligned phonemes. In *21st International Society for Music Information Retrieval (ISMIR)*, 2020. (Cited on pages 111 and 145.)

G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters. Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. *International Society for Music Information Retrieval (ISMIR),* Paris, France, 2018. (Cited on page 111.)

G. Meseguer-Brocal, R. Bittner, S. Durand, and B. Brost. Data cleansing with contrastive learning for vocal note event annotations. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 2020a. (Cited on page 58.)

G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters. Creating dali, a large dataset of synchronized audio, lyrics, and notes. *Transactions of the International Society for Music Information Retrieval (ISMIR)*, 3(1), 2020b. (Cited on page 111.)

H. Miao, G. Cheng, C. Gao, P. Zhang, and Y. Yan. Transformer-based online ctc/attention end-to-end speech recognition architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6084–6088. IEEE, 2020. (Cited on page 83.)

A. H. Michaely, X. Zhang, G. Simko, C. Parada, and P. Aleksic. Keyword spotting for google assistant using contextual speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 272–278. IEEE, 2017. (Cited on page 48.)

O. Migliore and N. Obin. At the interface of speech and music: A study of prosody and musical prosody in rap music. In *Speech Prosody*, 2018. (Cited on page 32.)

A.-r. Mohamed. *Deep Neural Network Acoustic Models for ASR.* PhD thesis, University of Toronto, 2014. (Cited on page 69.)

A.-r. Mohamed, G. E. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1):14–22, 2011. (Cited on page 69.)

M. Mongeau and D. Sankoff. Comparison of musical sequences. *Computers and the Humanities*, 24:161–175, 1990. (Cited on page 67.)

M. Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007a. (Cited on page 67.)

M. Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007b. (Cited on page 42.)

M. Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*, volume 5. Springer, 2015. (Cited on page 33.)

M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. In *International Society for Music Information Retrieval (ISMIR)*, volume 2005, page 6. Citeseer, 2005. (Cited on page 67.)

T. Nakamura, E. Nakamura, and S. Sagayama. Real-time audio-to-score alignment of music performances containing errors and arbitrary repeats and skips. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):329–339, 2015. (Cited on page 67.)

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48 (3):443–453, 1970. (Cited on pages 44, 49, 50, 66, and 153.)

P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–874, 2001. (Cited on page 50.)

Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19):4050, 2019. (Cited on page 41.)

Y. Ohishi, M. Goto, K. Itou, and K. Takeda. Discrimination between singing and speaking voices. In *Ninth European Conference on Speech Communication and Technology*, 2005. (Cited on page 40.)

D. W. Otter, J. R. Medina, and J. K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32 (2):604–624, 2020. (Cited on page 55.)

D. Pal, C. Arpnikanondt, S. Funilkul, and V. Varadarajan. User experience with smart voice assistants: the accent perspective. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2019. (Cited on page 42.)

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015. (Cited on page 109.)

M. Papakostas, E. Spyrou, T. Giannakopoulos, G. Siantikos, D. Sgouropoulos, P. Mylonas, and F. Makedon. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation*, 5(2):26, 2017. (Cited on page 42.)

H. Park, C. Kim, H. Son, S. Seo, and J.-H. Kim. Hybrid ctc-attention network-based end-to-end speech recognition system for korean language. *Journal of Web Engineering*, pages 265–284, 2022. (Cited on page 90.)

G. Peeters and F. Angulo. Ssm-net: feature learning for music structure analysis using a self-similarity-matrix based loss. *arXiv preprint arXiv:2211.08141*, 2022. (Cited on page 95.)

G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, 2011. (Cited on page 42.)

A. Petit. Le travail de la pop : écouter *1989* (2014) de Taylor Swift en musicologue. Colloque *Le Silence du mainstream,* Strasbourg, 2022. https://www.canalc2.tv/video/16316 (accessed February, 26th 2022). (Cited on pages 29, 140, and 152.)

K. Pisanski, A. Anikin, and D. Reby. Vocal size exaggeration may have contributed to the origins of vocalic complexity. *Philosophical Transactions of the Royal Society B*, 377(1841): 20200401, 2022. (Cited on page 134.)

P. Placeway and J. Lafferty. Cheating with imperfect transcripts. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 4, pages 2115–2118. IEEE, 1996. (Cited on pages 67 and 68.)

E. Ponsot, P. Arias, and J.-J. Aucouturier. Uncovering mental representations of smiled speech using reverse correlation. *The Journal of the Acoustical Society of America*, 143(1): EL19–EL24, 2018. (Cited on page 32.)

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011. (Cited on page 115.)

F. Poyatos. *Nonverbal Communication across Disciplines.* John Benjamins Publishing, 2002. (Cited on page 136.)

D. S. Prerau. Computer pattern recognition of printed music. In *Proceedings of the November 16-18, 1971, fall joint computer conference*, pages 153–162, 1972. (Cited on pages 48 and 153.)

H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, 2019. (Cited on page 55.)

L. R. Rabiner and B. Gold. Theory and application of digital signal processing. *Englewood Cliffs: Prentice-Hall*, 1975. (Cited on page 33.)

C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel. Mir_eval: A transparent implementation of common mir metrics. In *International Society for Music Information Retrieval (ISMIR)*, pages 367–372, 2014. (Cited on page 112.)

Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner. The musdb18 corpus for music separation, 2017. (Cited on page 137.)

E. Ramasso, M. Rombaut, and D. Pellerin. Forward-backward-viterbi procedures in the transferable belief model for state sequence analysis using belief functions. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 9th European Conference, ECSQARU 2007, Hammamet, Tunisia, October 31-November 2, 2007. Proceedings 9*, pages 405–417. Springer, 2007. (Cited on page 49.)

C. Raphael. Automatic segmentation of acoustic musical signals using hidden markov models. *IEEE transactions on pattern analysis and machine intelligence*, 21(4):360–370, 1999. (Cited on page 67.)

C. Raphael. Music plus one: A system for flexible and expressive musical accompaniment. *Proc. of the ICMC, 2001*, 2001. (Cited on page 50.)

C. Raphael. Aligning music audio with symbolic scores using a hybrid graphical model. *Machine learning*, 65(2-3):389–409, 2006. (Cited on page 68.)

L. Renault, A. Vaglio, and R. Hennequin. Singing language identification using a deep phonotactic approach. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 271–275. IEEE, 2021. (Cited on pages 71 and 111.)

G. Révész. *Introduction to the psychology of music.* University of Oklahoma Press, 1954. (Cited on page 37.)

F. Ridzuan and W. M. N. W. Zainon. A review on data cleansing methods for big data. *Procedia Computer Science*, 161:731–738, 2019. (Cited on page 58.)

F. Rigaud and M. Radenen. Singing voice melody transcription using deep neural networks. In *International Society for Music Information Retrieval (ISMIR)*, pages 737–743, 2016. (Cited on page 48.)

A. Röbel and X. Rodet. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *International Conference on Digital Audio Effects*, pages 30–35, 2005. (Cited on page 38.)

H. Rodriguez, P. Arias, and C. Canonne. Investigating the pragmatic effects of musical syntax through musical humor. In *15th International Symposium of Cognition, Logic, and Communication*, 2021. (Cited on page 32.)

A. Roebel and F. Bous. Neural vocoding for singing and speaking voices with the multi-band excited wavenet. *Information*, 13(3):103, 2022. (Cited on page 145.)

I. Rosenfelder, J. Fruehwald, K. Evanini, S. Seyfarth, K. Gorman, H. Prichard, and J. Yuan. Fave (forced alignment and vowel extraction) 1.1.3. web resource, http://dx.doi.org/10.5281/zenodo.9846, 2017. accessed 22 February 2023. (Cited on page 68.)

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. (Cited on pages 59 and 61.)

B. Russell. *The problems of philosophy (Chapter 10)*. OUP Oxford, 2001. (Cited on page 110.)

M. P. Ryynänen and A. P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008. (Cited on page 48.)

L. Salais, P. Arias, C. Le Moine, V. Rosi, Y. Teytaut, N. Obin, and A. Roebel. Production strategies of vocal attitudes. In *Interspeech 2022*, pages 4985–4989. ISCA, 2022. (Cited on pages 22, 31, 32, 133, and 152.)

M. Schedl, E. Gómez, J. Urbano, et al. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, 8(2-3):127–261, 2014. (Cited on page 42.)

D. Schmandt-Besserat. The evolution of writing. *International Encyclopedia of Social and Behavioral Sciences*, pages 1–15, 2014. (Cited on page 28.)

K. Schulze-Forster, C. S. Doire, G. Richard, and R. Badeau. Joint phoneme alignment and text-informed speech separation on highly corrupted speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7274–7278. IEEE, 2020. (Cited on pages 70 and 115.)

K. Schulze-Forster, C. S. Doire, G. Richard, and R. Badeau. Phoneme level lyrics alignment and text-informed singing voice separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2382–2395, 2021. (Cited on pages 71, 81, 83, 85, 90, 145, and 153.)

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. (Cited on page 61.)

J. Serra and E. Gómez. A cover song identification system based on sequences of tonal descriptors. *Music Information Retrieval Evaluation eXchange (MIREX)*, 46, 2007. (Cited on page 67.)

C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. (Cited on page 25.)

R. N. Shepard. Geometrical approximations to the structure of musical pitch. *Psychological review*, 89(4):305, 1982. (Cited on page 37.)

F. Simonetta, S. Ntalampiras, and F. Avanzini. Audio-to-score alignment using deep automatic music transcription. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2021. (Cited on page 50.)

B. Sisman, J. Yamagishi, S. King, and H. Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157, 2020. (Cited on page 41.)

Y. Soullard, C. Ruffino, and T. Paquet. Ctcmodel: a keras model for connectionist temporal classification. Research report, Université de Rouen Normandie, 2019. hal-02420358. (Cited on page 114.)

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. (Cited on page 60.)

S. Srivastava, A. Verma, and S. Sharma. Optical character recognition techniques: A review. In *2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–6. IEEE, 2022. (Cited on page 47.)

F. Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020. (Cited on page 47.)

K. N. Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000. (Cited on page 30.)

S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of America*, 8(3):185–190, 1937. (Cited on page 37.)

D. Stoller, S. Ewert, and S. Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *International Society for Music Information Retrieval (ISMIR)*, 2018. (Cited on pages 71 and 145.)

D. Stoller, S. Durand, and S. Ewert. End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–185. IEEE, 2019. (Cited on pages 71, 80, 85, 111, 145, and 153.)

B. H. Story and K. Bunton. Vowel space density as an indicator of speech performance. *The Journal of the Acoustical Society of America*, 141(5):EL458–EL464, 2017. (Cited on page 134.)

H. R. Strayer. From neumes to notes: The evolution of music notation. 2013. (Cited on page 28.)

J. Sundberg. Articulatory interpretation of the "singing formant". *The Journal of the Acoustical Society of America*, 55(4):838–844, 1974. (Cited on pages 31 and 136.)

J. Sundberg. *The Science of the Singing Voice*. Northern Illinois University Press, 1987. (Cited on page 138.)

J. Sundberg and T. D. Rossing. The science of singing voice, 1990. (Cited on pages 30 and 40.)

H. Tachibana, K. Uenoyama, and S. Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4788. IEEE, 2018. (Cited on page 99.)

J. Tao, Y. Kang, and A. Li. Prosody conversion from neutral speech to emotional speech. *IEEE transactions on Audio, Speech, and Language processing*, 14(4):1145–1154, 2006. (Cited on page 41.)

Y. Teytaut and A. Roebel. Phoneme-to-audio alignment with recurrent neural networks for speaking and singing voice. In *Proceedings of Interspeech 2021*, pages 61–65. International Speech Communication Association; ISCA, 2021. (Cited on pages 22, 83, 91, 92, and 121.)

Y. Teytaut, B. Bouvier, and A. Roebel. A study on constraining connectionist temporal classification for temporal audio alignment. In *Interspeech 2022*, pages 5015–5019. ISCA, 2022. (Cited on pages 22, 84, 91, 92, 96, and 99.)

Y. Teytaut, A. Petit, C. Chabot-Canet, and A. Roebel. A musicological pipeline for singing voice style analysis with neural voice processing and alignment. In *Journées d'Informatique Musicale (JIM 2023), Saint-Denis*, 2023. (Cited on pages 22, 135, and 139.)

S. S. Tirumala, S. R. Shahamiri, and R. Garhwal, Abhimanyu Singh andWang. Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, 90:250–271, 2017. (Cited on page 42.)

A. Torres, A. Cabada, and J. J. Nieto. An exact formula for the number of alignments between two dna sequences. *DNA Sequence*, 14(6):427–430, 2003. (Cited on page 65.)

C. Tralie and E. Dempsey. Exact, parallelizable dynamic time warping alignment with linear memory. *International Society for Music Information Retrieval (ISMIR)*, 2020. (Cited on page 144.)

A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. d'Alché-Buc. Multilingual lyrics-to-audio alignment. In *International Society for Music Information Retrieval (ISMIR)*, 2020a. (Cited on pages 71, 82, 85, 113, 115, and 153.)

A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. d'Alché-Buc. Audio-based detection of explicit content in music. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 526–530. IEEE, 2020b. (Cited on pages 71 and 145.)

A. Vaglio, R. Hennequin, M. Moussallam, and G. Richard. The words remain the same: Cover detection with lyrics transcription. In *22nd International Society for Music Information Retrieval Conference ISMIR 2021*, 2021. (Cited on page 71.)

E. Variani, McDermott Erik, Lahouel Kamel, Bacchiani Michiel, and Bagby Tom. Sampled connectionist temporal classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4959–4963, 2018. (Cited on page 91.)

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. (Cited on pages 61, 83, and 84.)

V. Velichko and N. Zagoruyko. Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2(3):223–234, 1970. (Cited on pages 48 and 66.)

B. Vercoe. The synthetic performer in the context of live performance. In *Proceedings of International Computer Music Conference*, pages 199–200, 1984. (Cited on page 67.)

E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006. (Cited on pages 93 and 137.)

E. Vincent, T. Virtanen, and S. Gannot. *Audio source separation and speech enhancement*. John Wiley & Sons, 2018. (Cited on page 93.)

T. K. Vintsyuk. Speech discrimination by dynamic programming. *Cybernetics*, 4(1):52–57, 1968. (Cited on page 66.)

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967. (Cited on pages 49 and 66.)

S. Vogel, H. Ney, and C. Tillmann. Hmm-based word alignment in statistical translation. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996. (Cited on page 67.)

W. von Kempelen. *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. Bei JV Degen, Wien, 1791. (Cited on page 41.)

A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, et al. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018. (Cited on page 55.)

B. Řepová, M. Zábrodskỳ, J. Plzak, D. Kalfert, J. Matoušek, and J. Betka. Text-to-speech synthesis as an alternative communication means after total laryngectomy. 2021. (Cited on page 42.)

M. Wagner. Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms. In *ICASSP'81. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 1156–1159. IEEE, 1981. (Cited on page 67.)

S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017. (Cited on pages 71 and 90.)

C. Weiss and G. Peeters. Training deep pitch-class representations with a multi-label ctc loss. In *International Society for Music Information Retrieval (ISMIR)*, 2021. (Cited on pages 122 and 145.)

A. Wichmann. The attitudinal effects of prosody, and how they relate to emotion. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000. (Cited on pages 32 and 133.)

Y. Wu and T. S. Huang. Vision-based gesture recognition: A review. *Gesture-Based Communication in Human-Computer Interaction: International GestureWorkshop, GW'99 Gif-sur-Yvette, France, March 17-19, 1999 Proceedings*, pages 103–115, 1999. (Cited on page 47.)

X. Xia, X. Song, F. Luan, J. Zheng, Z. Chen, and X. Ma. Discriminative feature selection for on-line signature verification. *Pattern Recognition*, 74:422–433, 2018. (Cited on page 50.)

A. Ycart, A. McLeod, E. Benetos, K. Yoshii, et al. Blending acoustic and language model predictions for automatic music transcription. 2019. (Cited on page 146.)

Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville. Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720*, 2017. (Cited on page 71.)

J. Zhu, C. Zhang, and D. Jurgens. Phone-to-audio alignment without text: A semi-supervised approach. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8167–8171. IEEE, 2022. (Cited on page 146.)

X. J. Zhu. Semi-supervised learning literature survey. 2005. (Cited on page 58.)

V. Zue, S. Seneff, and J. Glass. Speech database development at mit: Timit and beyond. *Speech communication*, 9(4):351–356, 1990. (Cited on pages 40 and 109.)