A musicological pipeline for singing voice style analysis with neural voice processing and alignment

Yann Teytaut, Antoine Petit, Céline Chabot-Canet, Axel Roebel

A MUSICOLOGICAL PIPELINE FOR SINGING VOICE STYLE ANALYSIS WITH NEURAL VOICE PROCESSING AND ALIGNMENT

Yann Teytaut
STMS (UMR 9912)
teytaut@ircam.fr

Antoine Petit
Passages XX-XXI
antoine.petit@univ-lyon2.fr

Céline Chabot-Canet
Passages XX-XXI
celine.chabot-canet@univ-lyon2.fr

Axel Roebel
STMS (UMR 9912)
roebel@ircam.fr

ABSTRACT

The study of singing style is of great interest both for expressive vocal synthesis and for the musicological analysis of vocal performances, inciting to a fruitful convergence between signal processing and musicology. However, for musicologists, these studies often come up against the absence of automatic analysis tools for voices recorded in a musical context, leading to long and tedious manual annotation work. This constraint imposes either to limit oneself to a restricted corpus, or to circumscribe one's study to experimental corpora of voices without instrumental accompaniment, thus depriving oneself of the unequalled interest that commercial recordings represent, as accomplished artistic works. This article introduces a new protocol using deep learning techniques to provide musicologists with powerful tools for the analysis of singing voices, opening up new perspectives through the automation of the different steps. We present a complete processing chain in support of musicological analysis, using neural models to isolate singing voice, predict its F0, and automatically align the syllables or notes to the audio (despite the musical accompaniment). The effectiveness of this approach is demonstrated by its practical application on two popular songs. These tools, developed in an ANR project, will soon be available to the scientific community.

Keywords: musicology, voice, singing style, popular music, voice alignment, voice processing, deep learning.

RÉSUMÉ

L'étude du style de chant revêt un fort intérêt tant pour la synthèse vocale expressive que pour l'analyse musicologique de performances vocales, incitant à une convergence fructueuse entre traitement du signal et musicologie. Cependant, pour les musicologues, ces études se heurtent souvent à l'absence d'outils d'analyse automatique de voix enregistrées en contexte musical, amenant à un long et fastidieux travail d'annotation manuelle. Cette contrainte impose soit de se limiter à un corpus restreint, soit de circonscrire son étude à des corpus expérimentaux de voix sans accompagnement instrumental, se privant alors de l'intérêt inégalé que représentent, comme œuvres artistiques abouties, les enregistrements commerciaux. Cet article introduit un protocole inédit utilisant des techniques d'apprentissage profond pour fournir aux musicologues des outils performants d'aide à l'analyse des voix chantées, ouvrant, par l'automatisation des différentes étapes, des perspectives nouvelles. Nous présentons une chaîne de traitements complète en support à l'analyse musicologique, exploitant des modèles neuronaux pour isoler la voix chantée, prédire sa F0 et automatiquement aligner les syllabes ou notes à l'audio (malgré l'accompagnement musical). L'efficacité de cette démarche est demontrée par son application pratique sur deux chansons populaires. Ces outils, développés dans le cadre d'un projet ANR, seront bientôt disponibles pour la communauté scientifique.

Mots-clefs: musicologie, voix, style de chant, musiques populaires, alignement de voix, traitement de la voix, apprentissage profond.

1. INTRODUCTION

Expressivity is one of the core elements that come into play when humans communicate with each other. Indeed, by exploiting all the prosodic resources of the spoken voice [27] – intonation, stress, paralinguistic effects, etc. – speakers are able to express a wide variety of emotions and social attitudes [43]. The singing voice, as a vector of communication, appropriates many of these codes. As a result, in the 2010s, some musicologists initiated the application of paralinguistics, phonostylistics or psycholinguistics [37, 30, 18] to vocal performance analysis [26, 6].

Between prosody, paralinguistic impregnations, rhetorical procedures and singing techniques (register, timbre, etc.), singers make use of a rich palette of vocal effects. This palette contains elements that:

— connect singers to specific groups, relating to a *generic style* through conventions allowing the identification with a "tribe" – *e.g.*, smoothing of registers and singing formant in classical singing, yodeling and twang in country music [34], belting in musicals [25], guttural voice in extreme metal [22], etc.;

— make singers unique, relating to a *personal style* – *e.g.*, particular vibrato, specific phrasing, intonative effects, timbre, manner of attacking, sustaining or ending notes, etc.

These characteristics can operate on a global scale (*e.g.*, the roughness of a voice) or appear locally, as emphasis. They can be difficult to define, however, for they often strike a subtle balance between antagonistic entities –

speech and song, harmonicity and noise, pure musicality and valorization of the text. This is all the more true in popular music, since its norms and conventions are looser than in classical music, due, in particular, to the absence of explicit theorization. The palette of vocal effects is potentially infinite, and even apparent mistakes can be admitted and deliberately used [8].

Although a well-defined artistic identity implies a certain amount of coherence, each performance is unique: choices made according to the needs of the song as well as the context enact a "*mouvance*" [29] characteristic of a music brought about by extemporization [4], the transience of the moment fortunately fixed by the recording, which also adds its own layer of complexity. To identify the style of an artist is thus not only to describe in detail such or such performance, but also to deduce its potentialities, its virtualities, to understand the general strategies, the processes at work, and the challenges that they support.

If the study and the interpretation of stylistic data resulting from the analysis of singing voice in a musical context can take the most diverse paths according to the disciplinary angle chosen (musicological, anthropological, philosophical, cultural, gender studies, etc.), a first stage is often essential: the description, in the most neutral and objective way possible, of the sonic materiality of the voice, as mediated by the recording and, possibly, additional studio processing (production effects).

The meticulous observation of the vocal phenomenon enables one to, for example, establish unsuspected parallels between vocal effects or techniques carrying very different meanings according to the generic, cultural and aesthetic contexts in which they appear. It is for these reasons that a growing branch of musicology is turning to the study of spectral representations of sound (sonograms), following the path traced by a few pioneers of the 1980s [11]. If we disregard identifiable production effects as such, and focus on purely vocal effects, the infinite number of observable phenomena can be abstracted to a limited number of acoustic parameters, pertaining to pitch, prosody, and quality (*i.e.*, timbre, see also [24]), which interact to create a specific *vocal delivery* [31].

The study of singing style [7], *i.e.*, the production strategies at play in a singer's performances, and the palette of delivery effects that defines their artistic identity, is of great interest to musicologists, but the modeling of singing style also has direct applications for the synthesis of a more expressive and natural-sounding singing voice [2]. This convergence of interests has led to a close collaboration between signal processing and musicology researchers, initiating an unprecedented situation of interaction.

For years, musicologists studying singing performance have been faced with the absence of tools for automating the processes of acoustic analysis and annotation of corpora of voices recorded in a musical context, leading to long and tedious work of manual annotation (transcription) and synchronization (alignment), either by ear or through visual spectral representations like sonograms.

If the expert listening of musicologists remains indispensable to supervise both transcription and alignment, a purely manual system has its limits. It is very time-consuming, meaning the corpus must either be small, or an experimental one, comprised of voices recorded without instrumental accompaniment, which is of little interest to musicologists interested in the *actual* musical works found in commercial recordings. It can also be error-prone and, in some cases, overly subjective. Thus, relying on automatic systems for transcription [19] and alignment [40] can lead to a considerable gain in time, and help in setting a common base for the musicological community.

In this context, deep learning has found numerous applications for voice-related tasks yet, to the best of the authors' knowledge, has not been directly dedicated to such musicological applications. In this work, taking advantage of the latest improvements in singing voice separation [10], F0 estimation [3], and voice alignment [16] algorithms based on deep learning, we introduce a complete pipeline which simplifies tedious analysis steps previously carried out by hand, and opens new perspectives through automation. With the intent to share these algorithms with the community, a companion website is under construction in the context of the ANR (French National Research Agency) project **A**nalysis and t**R**ansformation of **S**inging style (**ARS**) [1] .

More specifically, our proposed contributions are:

— A complete pipeline automatically extracting voice parameters upon state-of-the-art vocals separation, and aligning the audio with lyrics and notes;

— A coupling between expert transcriptions (lyrics, notes) and audio features, allowing (computational) analyses to be performed not only on symbolic (pitch, metric position) or acoustic (F0, syllable onset and length) data, but *across* both sets of data to bridge the semiotic divide identified by [21];

— Case studies illustrating the musicological interest of our unified pipeline on two popular songs from different languages and genres;

— A website set up to give access to the tools to the scientific community [2] .

2. RELATED WORKS

The current project relies on previous research dedicate to the analysis of singing style and its modeling for the synthesis of expressive singing, carried out within the framework of the ANR project **ChaNTeR** [3] , of which one can find a detailed example of application to Edith PIAF in the article [9]. As detailed in the complementary paper [2] and thesis [1], this project was aimed at incorporating expressivity, intimately related to singing style, in a concatenative synthesis system.

1 . https://ars.ircam.fr/

2 . https://passagesxx-xxi.univ-lyon2.fr/activites/projets-anr/projet-ars-analyse-et-transformation-du-style-de-chant-1

3 . **Cha**nt **N**umérique avec contrôle **Te**mps **R**éel, ANR-13-CORD-0011, 2011-2017

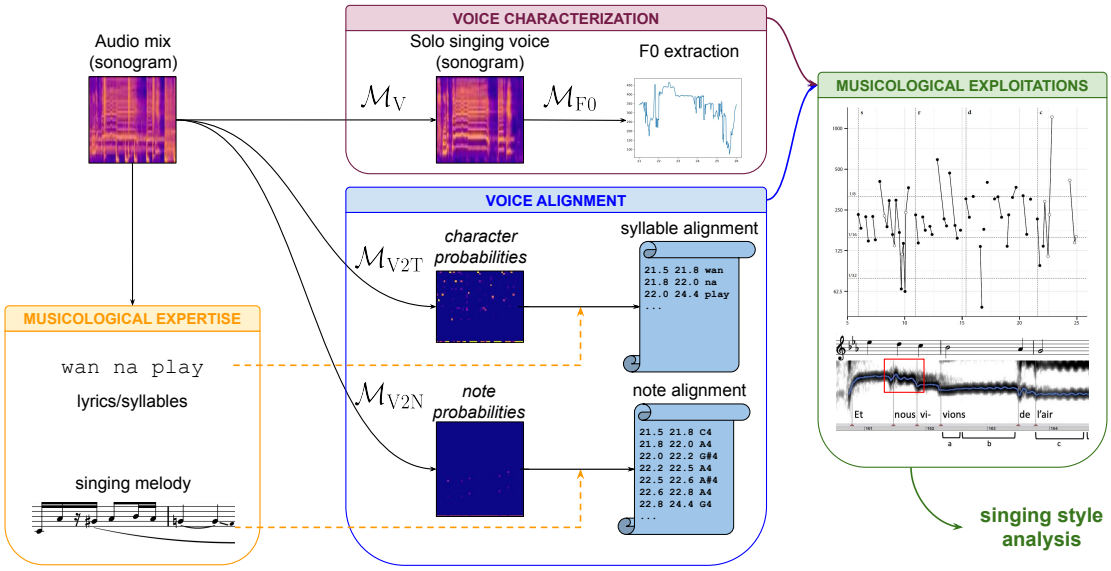

Figure 1. Overview of our complete analysis pipeline involving musicological expertise, deep learning models for the automation of voice characterization and alignment in order to help musicologists studying singing voice style.

Concretely, the synthesis of a given utterance was achieved by progressively connecting the relevant biphones (*i.e.*, two successive phonemes) of a pre-recorded singing dataset. The desired F0, derived from the score notes, was applied to each segment. Then, a musicologically-informed expressive control was developed to counteract the "unnatural" effect induced by the concatenations.

Based on extensive analyses of several performances of a representative corpus of singers illustrating several generic singing styles, a series of style-defining vocal effects were identified and described. The musicologist also established a series of local contexts to determine correlations between specific musical situations (pitch and duration of notes, position in the phrase and in the song structure, etc.) and the occurrence of the described effects. Decision trees were created to associate these contexts with each note in the score. Finally, global contexts were used to further transform the vocal parameters (*e.g.*, F0, intensity, etc.) in a consistent and expressive way.

The musicological analysis of the corpus required time-aligned lyrics and MIDI notes with respect to the audio. However, at that time, the presence of an instrumental accompaniment prevented the use of available systems, designed for *solo* singing voice. This difficulty could not be circumvented, as the desire to study iconic singers, such as Edith PIAF or Jacques BREL, required the use of commercial recordings, for which we did not have access to the vocal track independently from the full mix. As a result, the annotations and alignments were done entirely by hand. This process being very time-consuming, the choice was made to limit the corpus to 4 songs per artist studied, which offered sufficient data to obtain convincing results in the modeling of styles in a reasonable amount of time.

In this work, consequently, one of our key goals is to greatly simplify these processes and open the way for a complete and *easy-to-use* musicological analysis pipeline.

By exploiting recent deep learning models, F0 extraction is obtained automatically upon high-quality separation of the voice from the accompaniment, and lyrics and notes alignments are generated automatically even with background music present. These unprecedented possibilities open new perspectives for musicology, with the opportunity to work systematically on large corpora of recordings, considerably widening its field of investigation.

Because they contain many more instances of each effect and context, these large corpora allow for a more reliable identification of an artist's style, and facilitate, *e.g.*, the search for similarities or voice synthesis opportunities.

3. A PIPELINE FOR SINGING STYLE ANALYSIS

This section presents our proposed pipeline. As shown in Figure 1, it consists in four main categories with highly interdependent modules, namely *voice characterization*, *musicological expertise*, *voice alignment*, and *musicological exploitations*, which are further detailed.

Note that, although the modules are introduced in their logical order of operation, the pipeline offers musicologists a more varied set of workflows: the system is flexible enough to accommodate different perspectives and practices, as well as multiple musical ontologies (*i.e.*, kinds of musical works) – see subsection 3.4.

3.1. Voice characterization

The core of any singing or singer analysis system is the voice itself. In the context of commercial recordings, the presence of background music is a major problem for state-of-the-art parameter estimation algorithms and hinders a precise description of the singers' intonation or intensity contours. Thus, this first step aims to extract the singing voice from the background music.

3.1.1. Singing voice separation

The state of the art in source separation relies on deep learning techniques [33]. In our application, the separated vocals only serve for parameter estimation, such that a single-channel (mono), 16kHz extraction is sufficient. From the many proposed neural architectures, we chose to re-implement the network presented in [10], as it achieved state-of-the-art singing voice extraction quality with a comparatively small number of parameters.

This model, denoted as $\mathcal{M}_{\text{V}}$, has been trained using the publicly available MUSDB and CCMixter data sets, and a collection of internal data featuring solo singing voices and instrumental music including notably instruments not well covered in the public data sets. During training, the voice and music samples were randomly mixed and pitch-shifted following [12, 28]. The final model allows for efficient vocal signal separation with a very satisfying quality. It achieves an SDR of **9.2dB** for the vocals separated from the test set of the HQ version of the MUSDB [38] data set, which compares favourably with state-of-the-art performances [33]. In inference, separation is faster than real-time even when running on a CPU on a small laptop.

3.1.2. Parameter extraction

The isolated voice signal is then used to extract singing voice parameters. Currently, only the fundamental frequency (F0) and intensity contours are covered.

The F0 estimation algorithm is also a deep neural network $\mathcal{M}_{\text{F0}}$ that has been trained using a large data set of speech signals using an analysis/resynthesis procedure that allowed the creation of a perfect annotation of the target F0 contours [3]. Despite having been trained on clean speech only, this algorithm has recently shown very good performances for pathological voice signals as well [41]. The F0 estimation has been enhanced with a voice/unvoiced algorithm forcing the estimated F0 to zero for silent parts, or segments not dominated by a single voice.

The intensity estimation is performed by means of calculating the root mean average energy of the separated voice signal over short analysis windows.

As these algorithms are limited to solo vocals, future works may focus on their robustness to polyphonic music.

3.2. Musicological expert knowledge

Upon characterization of the solo singing voice, a typical musicological objective is to correlate voice features with others, such as transcriptions of relevant information, and emphasize their relationships.

Transcription is a task aiming at predicting a symbolic sequence from another data representation. In the context of voice analysis, two transcription tasks are ubiquitous, highlighting two modalities of utmost importance: text and meloody. A naive transcription may miss elements and subtleties that precisely describe the vocal performance – therefore, expert knowledge and supervision by musicologists are essential.

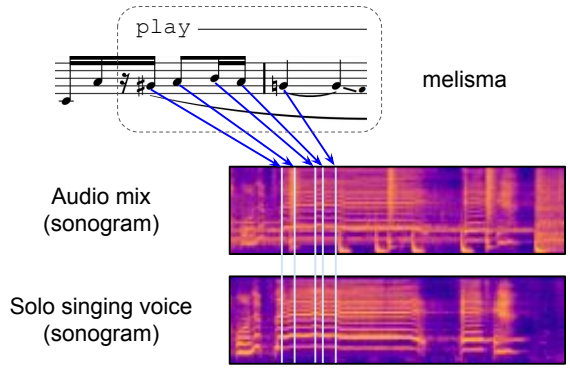

Figure 2. Illustration of the automatic analysis of a melisma: syllable-level alignment only predicts the word "play" on the full duration of this excerpt, without taking pitch variation into account – note-level alignment allows a deeper look into this gesture. See Taylor SWIFT case study – subsection 4.1.

3.2.1. Lyrics retrieval

For most vocal music today, lyrics are easily available online or via the album booklet, so that manual transcription is rarely required. However, instead of a succession of words, a *syllable*-level segmentation of the text is more relevant as singing notes are held on syllable vowels [39]. In practice, the lyrics rarely match the singing content perfectly due to additive onomatopoeia (*e.g.*, "yeah") or unpronounced utterances. Such local irregularities, fortunately, are not a problem for recent audio-to-text aligners, which can handle missing or additional syllables [16].

The role of the musicologists in lyrics retrieval is twofold: to ensure that the text is coherent and correctly written; and to explicitly adapt, whenever necessary, repeated syllables, missing entries, or onomatopoeia judged pertinent (*i.e.*, conveying meaningful interpretative aspects).

3.2.2. Melodic transcription

Melodic transcription consists in determining the notes performed by the singer in, *e.g.*, musical notation. In opposition to an F0 extraction, associated with *performance time*, a musical score is concerned with *symbolic time*.

A dedicated algorithm may be used as a first step, provided that it is robust and able to distinguish notes of the singing voice from accompaniment notes, to help in the process. In this paper, transcriptions were done entirely by musicologists, without recourse to such an algorithm.

A resulting score transcription can include one or several instances of *melisma*, *i.e.*, multiple notes sung on the same syllable (shown in musical notation by slurs). In this study, musicologists specifically rely on note-level alignment to complement the syllable alignment in such cases. An example is depicted in Figure 2. ***To the best of the authors' knowledge, it is the first option proposed to musicologists for dealing with the automatic analysis of melisma.***

3.3. Voice-to-symbols synchronization

With the audio and symbolic sequences (syllables and notes) at disposal, an *alignment* algorithm aims to associate each element in the sequences with a time in the audio, corresponding to its onset, a key step to further study the temporal aspects in singing performances.

3.3.1. CTC-based neural alignment

The alignment models are also deep neural networks, trained to minimize the Connectionist Temporal Classification (CTC) [20], a loss function assessing sequence-to-sequence prediction that does not need aligned training data. Given a spectral representation of the audio − with a theoretical time precision δt of 16ms in our setup −, a CTC-based model outputs, for each frame, a probability distribution over an alphabet $\mathcal{A}$ of L symbols plus a non-informative blank token. (See examples in Figure 1.) These per-frame probabilities can be used to force-align an audia with a sequence via a CTC variant of VITERBI's decoding. The architecture of our acoustic modes (deep CNN) and the decoding module are the same as [16].

Training is done with the large collection of English songs with roughly aligned words and notes from the DALI dataset [32]. The models have the great benefit of being applicable to all languages sharing the same alphabet $\mathcal{A}$ (although specialized, hence better, on English − see [16]) and across various musical genres irrespective of production date, meaning musicologists do *not* need to adjust parameters. This work precisely uses the same models on an American pop song from 2014 and a French chanson song from 1966. Finally, ***singing voice separation is not a mandatory step for running the alignment models***, which was a major technical deadlock imposing manual alignments in [9], so that their usage goes beyond our pipeline.

This approach does not outperform the state of the art [19] but is independent from any domain knowledge [16].

3.3.2. Aligning audio with syllables and notes

Let $\mathcal{M}_{\text{V2T}}$ denote the voice-to-text (V2T) aligner. The alphabet $\mathcal{A}$ contains all the basic latin characters (a, b, etc.), digits (0, 1, etc.), and a space token ø for separating successive syllables, hence $L = 37$. Although designed for word-level alignment, the model can also synchronize syllables, as words and syllables share the same alphabet. The only difference lies in the decoding step, as there are more spaces ø between syllables than words.

Let $\mathcal{M}_{\text{V2N}}$ denote the voice-to-note (V2N) aligner. We retrieved the F0 annotations in DALI and converted them into notes that range from C_1 to C_7 − this is particularly large for the human voice, but a manual inspection of outliers has not been pursued. The alphabet $\mathcal{A}$ contains all 12 semitones per octave and a silence token for long pauses (>500ms in DALI annotations), hence $L = 73$. ***It is, to the best of the authors' knowledge, the first time that an end-to-end, CTC-based model addresses note alignment*** − while note transcription was tackled [42].

	V2T	V2N
Playlist50 [16]†		
⌞ AAE (ms)	124.2 ± 58.5	232.0 ± 263.6
⌞ MAE (ms)	39.2 ± 14.1	105.1 ± 170.6
⌞ CER (%)	49.1 ± 14.2	39.7 ± 11.2
"Blank Space" (manual correction)		
⌞ AAE (ms)	13.5 ± 40.9	22.5 ± 63.0

† Metrics are averaged over all songs.

Table 1. Voice-to-text (V2T) and voice-to-note (V2N) alignment evaluations in terms of average (AAE) and median (MAE) alignment errors and character error rate (CER).

3.3.3. Alignment accuracy

We briefly assess the robustness of our models on the Playlist50 evaluation set, introduced in [16], in terms of alignment and transcription, as shown in the Table 1. Results show that it is more challenging to align notes than syllables, although overall mean errors remain below the commonly admitted 300ms perceptive threshold [13]. However, recognition of notes is better than lyrics − but it is known that a better transcription does not systematically imply a better alignment in a CTC framework [40]. Interestingly, in our "Blank Space" study, exploiting both alignments (see 4.1), the manual syllable corrections are *below* the theoretical precision δt while note alignment, although once again less stable, is very much acceptable.

3.4. Musicological exploitations

Finally, our proposed protocol allows further musicological studies while offering great flexibility.

The system only expects that symbolic data (text, notes) can be related to a recording, meaning it works equally well whether this data is transcribed from the recording or exists prior to it, as in *written* vocal music (*e.g.*, art songs), the study of which is also possible with this pipeline.

Because most modules output files directly and independently from one another, musicologists can freely decide which to use and how. One approach, which has been favored by one of the authors, is to use the data from each module to repeatedly refine the score transcription, thus enacting a sort of back and forth between symbolic and acoustic data. Also, because the data is not tied to a single working environment, it is available for computational analysis, as shown in subsection 4.1, which uses a series of scripts written by the musicologist in the R programming language.

The time markers from the alignment files can be be manually corrected using visualization software, *e.g.*, Sonic Visualiser [4] or RX [5] − a process much less tedious than starting from scratch. As for voice parameters, F0 estimation curves require close to no manual corrections.

4. https://www.sonicvisualiser.org
5. https://www.izotope.com/en/products/rx.html

4. MUSICOLOGICAL CASE STUDIES: TAYLOR SWIFT & CHARLES AZNAVOUR

Having introduced our general pipeline, a demonstration of the whole musicological protocol is proposed via two case studies, first on Taylor SWIFT's 2014 hit single "Blank Space" [6] and then on Charles AZNAVOUR's "La Bohème" from 1966 [7]. In the first study (SWIFT), temporal data from the alignments is used to perform fine-grained rhythmic analyses and to investigate the structural role played by articulation and micro-rhythm. In the second one (AZNAVOUR), the fundamental frequency estimation is used to study vocal phrasing and rhetorical effects involving intonation.

4.1. Taylor SWIFT's "Blank Space"

This song was selected for several reasons: (1) it is (at time of writing) SWIFT's second-best charting song, having stayed seven non-consecutive weeks at the top of the Billboard Hot 100, and its analysis may contribute to a more thorough understanding of what makes a successful song; (2) it, and SWIFT's songs more generally, have not yet been the object of much, if any, musicological attention (although see [36]); and (3) it is in standard compound AABA form, but it exhibits somewhat intricate patterning at lower levels of organization, with verse and chorus each articulating two iterations of an **srdc** [8] structure [17] in part through shifts in vocal delivery. It thus presents a prime example of the kind of analytical work afforded to musicologists by the pipeline. The following case study focuses on the first half of the first verse (eight bars, from `5:30` to `25:30`), a score transcription of which is given in Figure 3.

This excerpt can be divided into four parts: two fairly similar segments (**sr**, from `5:30` to `15:30`) and a contrasting passage leading to a concluding gesture (**dc**, from `15:30` to `25:30`). Following [23], we may wish to understand which criteria elicit such a segmentation, especially when it appears so self-evident.

A typical analysis, focusing mostly on pitch, would highlight the close resemblance between **s** and **r**, which share a series of two-syllables phrases [9] on F_4 followed by a syncopated descending step-wise motion from A_4 back to F_4 (with an additional leap from G_4 to D_4 in **s**). They would then be contrasted with the series of short phrases that make up **d**, comprised exclusively of large leaps which also instantiates a hierarchy divorce [35] with the underlying harmony (C and A over a B♭ chord, and F and A over a C chord), and the long melisma on "play?" that defines **c**.

6 . Words and music by Taylor SWIFT, Max MARTIN et SHELLBACK. Reference recording: *1989*, Big Machine, 2014.

7 . Words by Jean PLANTE, music by Charles AZNAVOUR. From the operetta *Monsieur Carnaval*, 1965. Reference recording: *La Bohème*, Barclay, 1966.

8 . **S**tatement, **r**estatement/**r**esponse, **d**eparture, **c**losure.

9 . Following convention, we equate "phrase" with "breath group". Comma-like symbols in the transcription indicate the points at which SWIFT takes a breath.

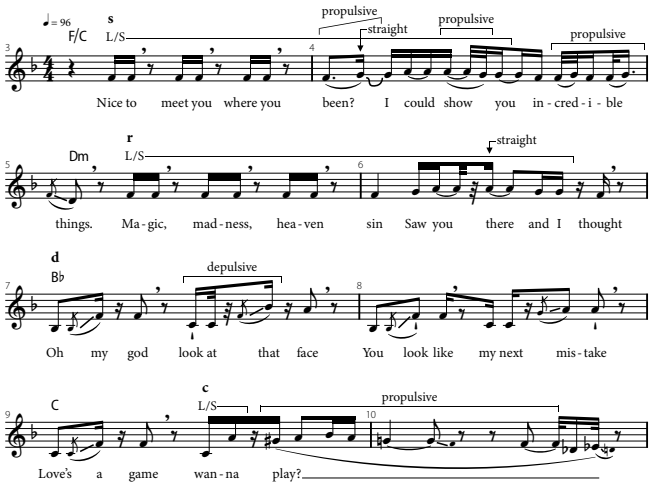

Figure 3. Taylor SWIFT, "Blank Space", measures 3–10. [Transcription done by Antoine PETIT.]

Although such analysis is relevant, we believe that articulation and micro-rhythm play as much, if not more, of a role in shaping the form of the excerpt − dimensions that our proposed pipeline precisely allows us to investigate.

4.1.1. Application of the pipeline to the song

Syllables retrieved from a first score transcription were synchronized with the audio by infering with the $\mathcal{M}_{\text{V2T}}$ model. The syllable alignment was then processed with an R script to delete the end marker of syllables *not* followed by a rest in the transcription (*i.e.*, when the articulation was heard as *legato*, meaning the end of the n syllable and the onset of the $n+1$ syllable can be taken to be identical). The same script was also used to generate an additional marker for every syllable, which is the mean of the onset and end time; this marker is meant to approximate the perceptual center (P-center) of the syllable (*i.e.*, the moment it is *heard* as beginning, as opposed to its acoustic onset − see [14] and [15] for inherent limits on P-center representation as singular time-points). The processed syllable alignment − 121 markers for 50 syllables − was then imported into RX for manual correction: the onset and (when present) end markers were corrected visually using a sonogram (with aural checking when necessary); the P-centers were systematically checked aurally by converting them to clicks using Sonic Visualiser.

The whole process was completed in roughly over an hour (*i.e.*, two markers per minute): an acceptable benchmark, given that roughly half of the markers require attentive listening upon correction, which can be expected to decrease with further experience, and a far cry from the countless hours previously required for this task.

As previously mentioned, the corrected syllable alignment was then used to revise the transcription. In particular, it allowed for a more accurate transcription of pitches when paired with the automatically-extracted F0. The resulting string of MIDI notes was then synchronized with

	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 2. Non-*legato* notes by subsection in Taylor SWIFT's "Blank Space", measures 3–10.

the audio by infering with the $\mathcal{M}_{\text{V2N}}$ model, and the note alignment was processed with another R script to delete all notes *not* part of a melisma, as well as the first note of every melisma (which had already been aligned using the syllables). The 10 remaining notes were again imported in RX for manual correction and, because we consider them to represent P-centers, aurally checked afterwards [10].

4.1.2. Musicological exploitation

With this timing data at hand − onset, P-center and end of every note in the excerpt − we are now able to propose an analysis of 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 $n+1$ note with the end time of the n note. Table 2 displays the number, proportion, and mean duration (with standard deviation) 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, at 304ms (almost half of a beat at 96BPM). Granted, this arch-like progression seems fairly obvious upon listening (especially when it has been explicitly pointed out beforehand), but we may very well have missed it had we not been able to gather accurate timing data.

Articulation is only part of the story, however. Figure 4 maps the duration of every note (*i.e.*, the difference between the P-center of the $n+1$ note and that of the n note, to which is subtracted the length of the intervening silence, if there is one) to its P-center, with *legato* articulation shown with connecting lines, and the internal notes of each melisma (those aligned with the $\mathcal{M}_{\text{V2N}}$ model) 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 (the three most frequent symbolic durations in the transcription) at 96BPM.

10 . Gliding intonations, such as those in **d**, are heard as a single gesture and are thus not considered as instances of melisma (this hearing is reflected in the transcription by smaller notes, approximating the initial or final pitch of the glide, followed by glissando lines).

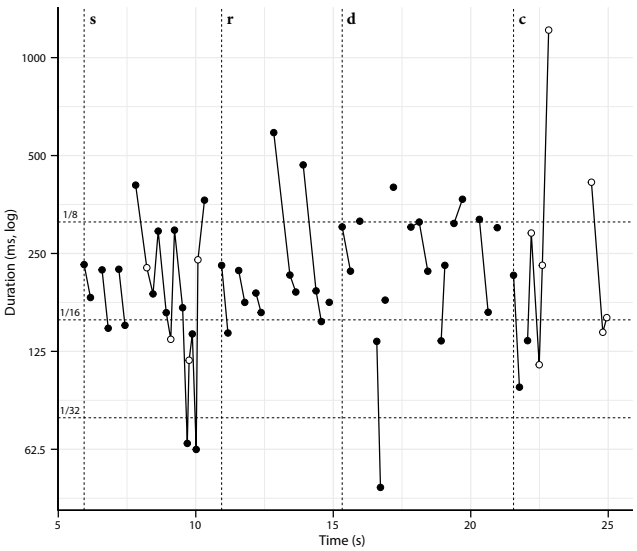

Figure 4. Note durations (log-scale) in Taylor SWIFT's "Blank Space", measures 3–10.

	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 3. Micro-rhythmic displacements by subsection in Taylor SWIFT's "Blank Space", measures 3–10.

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 [5] (*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.

We can thus highlight 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.

These local styles also share a number of characteristics, among which propulsive tendencies [5] in melismatic passages (*i.e.*, the notes are shorter than projected) [11] and lengthened notes when followed by a silence (both of these can be observed on Figure 4), 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 3.

11 . 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".

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', and linked to, for example, the lyrics (do the L/S sixteenth notes connote out-of-breathness? do the few straight sixteenth notes constitute an early "mask-off" moment for the hysterical character portrayed by SWIFT in "Blank Space"? etc.).

4.2. Charles AZNAVOUR's "La Bohème"

We now turn to a second example of musicological exploitation, on a fundamentally different repertoire that highlights another facet of the possible uses of the pipeline: French chanson, illustrated by an emblematic track from an equally emblematic artist: "La Bohème", as sung by Charles AZNAVOUR in 1966. In addition to being in another language, the traditional French chanson aesthetic is markedly different from that of Anglo-Saxon pop music: the text's primacy, and its vocal enhancement, are distinctive traits that differentiate it from almost all other genres.

While articulation and micro-rhythm in the context of a strict pulse played a central role in defining Taylor SWIFT's style, French chanson also emphasizes other dimensions, such as paralinguistic effects involving the fundamental frequency. With this song's nostalgic tone exploiting pathos as its main tool for seducing the listener, the rhetorical use of emphatic or euphemistic effects will be the focus of the following case study, without, however, developing all their potentialities (rhythmic placement, for example, will not be touched upon very much).

4.2.1. Application of the pipeline to the song

The first two steps in the pipeline – voice separation and F0 estimation – give particularly good results with this song. The separated voice file provides the musicologist with a working support of sufficient quality for both aural and visual analysis (including that of timbre, possibly the dimension most affected by the source separation process, but of which both the harmonic and noisy components are preserved). The quality of the source separation makes it possible to obtain an accurate fundamental frequency curve with the $\mathcal{M}_{\text{F0}}$ model that does not require manual corrections beyond the occasional removing of the curve on silences and unvoiced consonants, whereas manual annotation as it was previously performed on a music file with accompaniment required between 30 minutes and 1 hour of work per song. Moreover, this automatically-extracted F0 is more precise than manually tracing the curve on the sonogram could ever be; it is thus an analytically reliable time saving method. The marker file resulting from the voice-to-text alignment allows the musicologist to instantaneously relate the effects heard, and observed on the sonogram, to the lyrics/syllables.

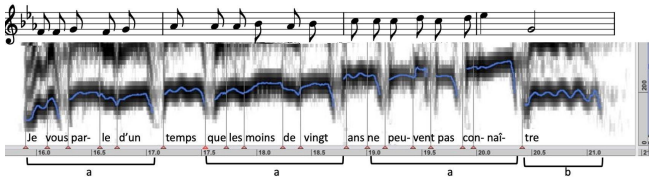

Figure 5. Verse 1, line 1 from Charles AZNAVOUR's "La Bohème". Score transcription, sonogram, F0 and text alignment (a = fast flow, intonational instability, and absence of vibrato; b = sustained note with vibrato on the last syllable of the phrase).

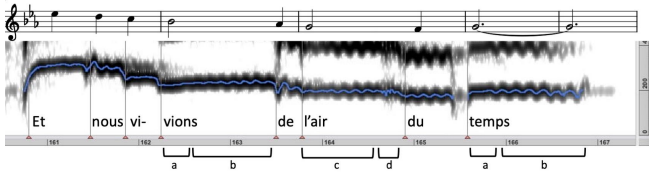

Figure 6. Chorus 3, line 4 from Charles AZNAVOUR's "La Bohème". Score transcription, sonogram, F0 and text alignment (a = beginning of a note without vibrato; b = arrival of the vibrato; c = vibrato on the note's whole duration; d = rolled "r").

4.2.2. Musicological exploitation

AZNAVOUR's style is characterized by the coexistence of two antagonistic, yet complementary, modes of delivery: (1) a narrative everyday-speech-like phrasing defined by a fluctuating intonation, rhythmic irregularity, and a lack of vibrato; and (2) a lyrical phrasing marked by sustained notes with vibrato, stable intonation, and rhythmic precision. The first one is shown in Figure 5, with its constantly rising and falling intonation over step-wise motion paired with a fast flow of eighth notes, and is found in verses. The second one appears in verses as well, on the last syllables of phrases, but is most characteristic of choruses, where rolled "r"s further pull AZNAVOUR's delivery away from the spoken voice and towards the singing voice. See Figure 6: the pitch remains stable throughout the notes' duration, while vibrato, which is under very fine control, is brought in gradually on the most meaningful words – an emphatic effect. Consonants are articulated and distinct, but are somewhat temporally and dynamically euphemized, giving the choruses a resolutely vowel-like character.

While this prosody-based formal structure is typical of the French chanson genre, it takes on a particular significance with AZNAVOUR. An important axis of his performance strategy is the pairing of profound musicality (expressed through intonative and rhythmic precision) with an apparent economy of means – a sobriety associated first and foremost with his well-known veiled tone, which he knew how to make great use of in the service of pathos, evoking the tragedy of the everyday life closely related to his songs' themes.

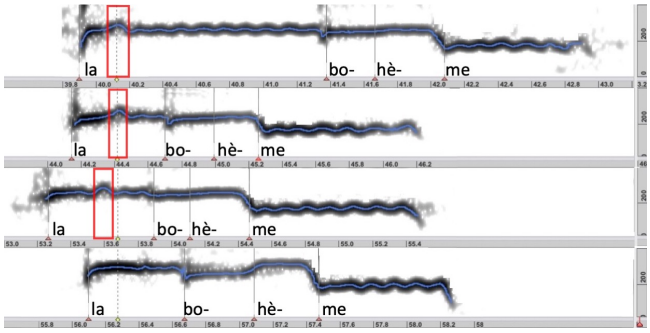

Figure 7. Synchronization of the 4 iterations of the "la bohème"-motif from the first chorus of Charles AZNAVOUR's "La Bohème". Sonogram, F0 curve and text alignment. The red frame highlights the position of the trill.

Another characteristic element of AZNAVOUR's singing style is the presence of vocal ornamentation, more specifically trills and appoggiaturas. "La Bohème" contains, to varying degrees of subtlety, about 24 such effects, which are found mostly in the lyrical parts: choruses and last lines of verses. One notable use of the trill in this song is to create variations throughout the 16 iterations of the title-phrase "la bohème", which appears 4 times par chorus on a single intonative scheme (transposed to different starting pitches) and rhythmic formula (three quarter notes followed by a half-note).

Figure 7 shows the first 4 iterations of the "la bohème"-motif (*i.e.*, those of the first chorus), synchronized in order to visualize both the distribution of the trill within the iterations and its rhythmic position. The synchronization was done not according to the onset of the first syllable, but to the rhythmic pulse of the beat as it materializes in the instrumental accompaniment, allowing the consideration of possible agogic (*i.e.*, micro-rhythmic) shifts.

5. CONCLUSION

In this paper, we introduced a complete pipeline for the musicological analysis of singing voice style. From a technical perspective, deep learning models were used for singing voice extraction from background music, voice parameter (F0) estimation, and 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 the two concrete case studies demonstrate: (1) text and note alignments allowed investigating articulation and micro-rhythm in a Taylor SWIFT song (American pop, 2014); and (2) F0 curves were exploited to highlight vocal phrasing and rhetorical effects involving intonation in a Charles AZNAVOUR song (French chanson, 1966). More generally, this work paves the way 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 the form of a web interface.

6. ACKNOWLEDGEMENTS

This work has been funded by the ANR (French National Research Agency) project ANR-19-CE38-0001-01 **ARS** (**A**nalysis and t**R**ansformation of **S**inging style).

7. REFERENCES

[1] Ardaillon, L. *Synthesis and expressive transformation of singing voice*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2017.

[2] Ardaillon, L., Chabot-Canet, C., and Roebel, A. "Expressive control of singing voice synthesis using musical contexts and a parametric F0 model". In *Proc. Interspeech 2016*, pages 1250–1254, 2016.

[3] Ardaillon, L. and Roebel, A. "Fully-convolutional network for pitch estimation of speech signals". In *Proc. Insterspeech 2019*, pages 2005–2009, 2019.

[4] Caporaletti, V. *I processi improvvisativi nella musica: un approccio globale*. Libreria Musicale Italiana, 2005.

[5] Caporaletti, V. *Swing e Groove: sui fondamenti estetici delle musiche audiotattili*. Libreria Italiana Musicale, 2014.

[6] Chabot-Canet, C. *Léo Ferré: une voix et un phrasé emblématiques*. L'Harmattan, 2008.

[7] Chabot-Canet, C. "L'analyse spectrale au fondement d'une rhétorique des styles interprétatifs dans la chanson française". *Volume!*, 16(2)/17(1): 29–47, 2020.

[8] Chabot-Canet, C. "L'interprétation de Barbara: l'expressivité de l'infime dans la performance vocale". In Bost, S. and Douzou, C. (ed.), *Barbara en scène(s)*, pages 105–130. Presses Universitaires de Provence, 2022.

[9] Chabot-Canet, C., Ardaillon, L., and Roebel, A. "Analyse du style vocal et modélisation pour la synthèse de chant expressif: l'exemple d'Édith Piaf". *Volume!*, 16(2)/17(1): 63–85, 2020.

[10] Choi, W., Kim, M., Chung, J., Lee, D., and Jung, S. "Investigating U-Nets with various intermediate blocks for spectrogram-based singing voice separation". *arXiv preprint arXiv:1912.02591*, 2019.

[11] Cogan, R. *New Images of Musical Sound*. Harvard University Press, 1984.

[12] Cohen-Hadria, A., Roebel, A., and Peeters, G. "Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation". In *EUSIPCO*, pages 1–5. IEEE, 2019.

[13] Cont, A., Schwarz, D., Schnell, N., and Raphael, C. "Evaluation of real-time audio-to-score alignment". In *ISMIR*, 2007.

[14] Danielsen, A., Nymoen, K., Anderson, E., Câmara, G. S., Langerød, M. T., Thompson, M. R., and London, J. "Where is the beat in that note? Effects of

attack, duration, and frequency on the perceived timing of musical and quasi-musical sounds". *Journal of Experimental Psychology: Human Perception and Performance*, 45(3): 402–418, 2019.

[15] Danielsen, A., Nymoen, K., Langerød, M. T., Jacobsen, E., Johansson, M., and London, J. "Sounds familiar(?): expertise with specific musical genres modulates timing perception and micro-level synchronization to auditory stimuli". *Attention, Perception, & Psychophysics*, pages 1–17, 2021.

[16] Doras, G., Teytaut, Y., and Roebel, A. "A linear memory CTC-based algorithm for text-to-voice alignment of very long audio recordings". *Applied Sciences*, 13(3): 1854, 2023.

[17] Everett, W. *The Foundations of Rock: From "Blue Suede Shoes" to "Suite: Judy Blue Eyes"*. Oxford University Press, 2009.

[18] Fónagy, I. *La Vive Voix: essais de psychophonétique*. Payot, 1983.

[19] Gao, X., Gupta, C., and Li, H. "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.

[20] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. "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.

[21] Guillotel-Nothmann, C. "Les signes musicaux et leur étude par l'informatique: le statut épistémologique du numérique dans l'appréhension du sens et de la signification en musique". *Revue musicale OICRM*, 6(2): 45–72, 2020.

[22] Hainaut, B. *Le style black metal*. Aedam Musicae, 2017.

[23] Hanninen, D. A. *A Theory of Music Analysis: On Segmentation and Associative Organization*. University of Rochester Press, 2012.

[24] Heidemann, K. "A system for describing vocal timbre in popular song". *Music Theory Online*, 22(1), 2016.

[25] Henrich Bernardoni, N. "La voix timbrée dans les chansons: considérations physiologiques et acoustiques". *Volume!*, 16(2)/17(1): 49–61, 2020.

[26] Lacasse, S. "The phonographic voice: paralinguistic features and phonographic staging in popular music singing". In Bayley, A. (ed.), *Recorded Music: Performance, Culture and Technology*, pages 225–251. Cambridge University Press, 2010.

[27] Lacheret-Dujour, A. and Beaugendre, F. *La Prosodie du français*. CNRS, 1999.

[28] Lancaster, E. P. and Souviraà-Labastie, N. "A frugal approach to music source separation", 2020.

[29] Le Vot, G. *Poétique du rock: oralité, voix et tumultes*. Minerve, 2017.

[30] Léon, P. *Précis de phonostylistique: parole et expressivité*. Nathan Université, 1993.

[31] Malawey, V. *A Blaze of Light in Every Word: Analyzing the Popular Singing Voice*. Oxford University Press, 2020.

[32] Meseguer-Brocal, G., Cohen-Hadria, A., and Peeters, G. "Creating DALI, a large dataset of synchronized audio, lyrics, and notes". *Transactions of the International Society for Music Information Retrieval*, 3(1): 55–67, 2020.

[33] Mitsufuji, Y., Fabbro, G., Uhlich, S., Stöter, F.-R., et al. "Music Demixing Challenge 2021". *Front. Signal Process.*, 1, 2022.

[34] Neal, J. "The twang factor in country music". In Fink, R., Latour, M., and Wallmark, Z. (ed.), *The Relentless Pursuit of Tone: Timbre in Popular Music*, pages 43–64. Oxford University Press, 2018.

[35] Nobile, D. "Counterpoint in rock music: unpacking the 'melodic-harmonic divorce"'. *Music Theory Spectrum*, 37(2): 189–203, 2015.

[36] Petit, A. "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).

[37] Poyatos, F. *Nonverbal Communication across Disciplines*. John Benjamins Publishing, 2002.

[38] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S., and Bittner, R. "The MUSDB18 corpus for music separation", 2017.

[39] Sundberg, J. *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

[40] Teytaut, Y., Bouvier, B., and Roebel, A. "A study on constraining Connectionist Temporal Classification for temporal audio alignment". *Proc. Interspeech 2022*, pages 5015–5019, 2022.

[41] Vaysse, R., Astésano, C., and Farinas, J. "Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech". *J. Acoust. Soc. Am.*, 152(5): 3091–3101, 2022.

[42] Weiß, C. and Peeters, G. "Learning multi-pitch estimation from weakly aligned score-audio pairs using a multi-label CTC loss". In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 121–125. IEEE, 2021.

[43] Wichmann, A. "The attitudinal effects of prosody, and how they relate to emotion". In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.