Intra-Sentential Speaking Rate Control in Neural Text-To-Speech for Automatic Dubbing

Mayank Sharma, Yogesh Virkar, Marcello Federico, Roberto Barra-Chicote, Robert Enyedi

Amazon

{mysharm, yvvirkar, marcfede, rchicote, renyedi}@amazon.com

Abstract

Automatically dubbed speech of a video involves: (i) segmenting the target sentences into phrases to reflect the speech-pause arrangement used by the original speaker, and (ii) adjusting the speaking rate of the synthetic voice at the phrase-level to match the exact timing of each corresponding source phrase. In this work, we investigate a post-segmentation approach to control the speaking rate of neural Text-to-Speech (TTS) at the phrase-level after generating the entire sentence. Our post-segmentation method relies on the attention matrix generated by the context generation step to perform a force-alignment over pause markers inserted in the input text. We show that: (i) our approach can be more accurate than applying an off-the-shelf forced aligner, and (ii) post-segmentation method permits generation more fluent speech than pre-segmentation approach described in [1].

Index Terms: Neural Text-to-Speech, Automatic Dubbing.

1. Introduction

Automatic Dubbing (AD) is the task of automatically localizing spoken content of a video document with speech in a different language, while preserving as much as possible the user experience. AD extends the task of speech translation [2, 3, 4, 5] in significant ways. In speech translation, a speech utterance in the source language is recognized, translated (and possibly synthesized) in the target language. Major additional requirements of AD are synchronization and transfer of paralinguistic aspects of the original utterance, like timbre, prosody and emotion. Regarding the synchrony requirements, generated speech should match the temporal speech-pause sequence of the original audio (isochrony), the articulatory movements of the mouth if it is visible in close-up scenes (lip synchrony), and possible body movements associated with the uttered words (kinesic synchrony) [6]. Isochrony is the most important form of synchronization [6] and similar to previous works [7, 1, 8] is the objective of this paper.

Isochrony implies that the translation of a dialogue line must be voiced by fitting precise temporal constraints, both at the phrase and pause levels. From the AD perspective, isochrony requires the machine translation (MT) systems generating translations of a specific length [9, 10] and text-to-speech (TTS) systems providing fine control of the speaking rate at the intra-sentential level, without compromising naturalness.

In this paper we focus on the TTS component, by discussing alternative ways to integrate speaking rate control in neural TTS as implemented in our AD research architecture [1]. In particular, we start from the naive method of first generating speech at the phrase level, then adjusting the duration of each phrase and pause, and finally collating all segments. However, despite its simplicity, this method does not guarantee good prosody across pauses. Hence, we consider alternative methods to generate speech of the entire sentence, segment it on the pauses, and then adjust the timing. We show that the most elegant and accurate solution is achieved by segmenting speech with the attention weights produced by the TTS generative model. Finally, we measure the quality of speech generated with the naive method and with the integrated method on a set of sentences in which we apply sub-sentential timing constraints typically of a dubbing scenario.

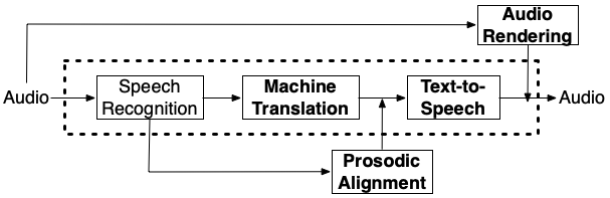

Figure 1: *Speech-to-speech translation pipeline (dotted box) with enhancements to perform automatic dubbing (in bold).*

Our paper is organized as follows. First, we describe our automatic dubbing architecture, by providing details about how input for TTS is generated and our basic neural TTS model. Next, we focus on the problem of controlling the TTS speaking rate at the intra-sentential level, by presenting the current solution and the new proposed approach, which implies segmenting the TTS output on pauses. Finally, we present experimental results with three TTS languages (French, German and Spanish) and three voices each, comparing accuracy of two segmentation methods and voice quality obtained by controlling the TTS speaking rate with the current and with the new method.

2. Automatic Dubbing

We build on the automatic dubbing architecture presented in [1, 8], and described in Figure 1, that extends a speech-to-speech translation [2, 3, 4] pipeline with: neural machine translation (MT) robust to ASR errors and able to control verbosity of the output [11, 10, 12]; prosodic alignment (PA) [7] which addresses phrase-level synchronization of the MT output by leveraging the force-aligned source transcript; neural text-to-speech (TTS) [13, 14, 15] with precise duration control; and, finally, audio rendering that enriches TTS output with the original background noise (extracted via audio source separation with deep U-Nets [16, 17]) and reverberation, estimated from the original audio [18, 19].

Automatic dubbing requires some enhancements to the conventional Neural TTS architecture, like the adjustment of the phrasing in the text input, and the incorporation of a precise control of the output length that matches the phrasing constraints imposed by the clip in the source language.

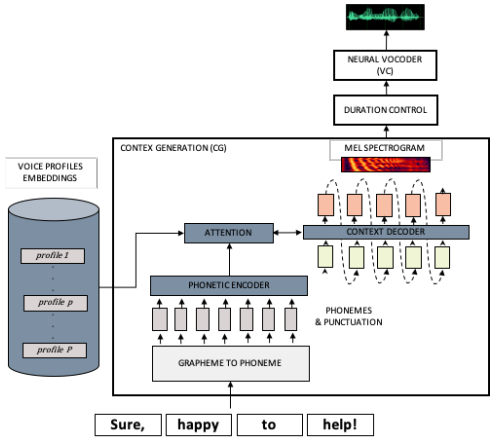

Figure 2: *Main modules of our neural TTS architecture: the context generation (CG), the duration control, and the neural vocoder (VC).*

2.1. TTS architecture

A state-of-the-art neural TTS system (Figure 2) consists of two modules: a Context Generation module (CG), which generates a context sequence from the input text and a Neural Vocoder module (VC), which converts the context sequence into a speech waveform. The first one is an attention-based sequence-to-sequence network [13, 14] that predicts a Mel-spectrum given an input text. A grapheme-to-phoneme module converts the sequence of words into a sequence of phonemes plus augmented features like punctuation marks and prosody related features derived from the text (e.g. lexical stress). For the Context Generation module, we trained multi-speaker models for each language. As shown in [14] the combination of recordings from several speakers increases the phonetic coverage and increases the robustness of the module. This approach enables scaling to the amount of target voices that a use-case as dubbing would require, without the need of large amounts of recordings per speaker. We use the Universal Neural Vocoder introduced in [15], pre-trained with 2000 utterances per each of the 74 voices from a proprietary database.
We trained a multi-speaker model for each of the three target languages: French, German and Spanish. Each training set involved a primary speaker with more that 20 hours of recordings, and at least two secondary speakers with around 10 hours of recordings individually. All the multi-speaker models consist of both female and male voices.

2.2. TTS Input

A schematic representation of how input for TTS is generated is shown in Figure 3. Given the transcript of a source sentence provided with time-stamps and position of significant pauses ($\geq$ 150 ms), a translation is generated of length close to the input [1] [9, 10]; then, *prosodic alignment* [7, 1, 8] is performed to place linguistically plausible pauses in the translation and delimit phrases matching the duration of the corresponding phrases in the source sentence.

[1] Measured in number of characters or syllables.

3. TTS Speaking Rate Control

Given the input segmentation and the timing of each phrase and pause, the goal of TTS is to produce a fluent output. This is achieved by controlling the speaking rate of each phrase and the duration of each pause, so that the given timing constraints are matched. Two possible approaches to produce fluent output are discussed in the following subsections.

3.1. Pre-Segmentation Approach

The approach implemented in [1] is depicted in Figure 3. Each phrase with corresponding target timing is passed to the context generator module (CG) which generates the Mel spectrum at the normal speaking rate. Prior to running the Neural Vocoder (VC), each Mel spectrum is linearly morphed along time to reach the target duration using a spline interpolation software. Finally, the resulting speech signals are concatenated with silence signals of the duration of the pauses in the source sentence. We empirically observe that this approach produces better quality speech than traditional time-stretching of the speech signal.

While this method is straightforward to implement, it has two significant drawbacks. First, TTS quality does often degrade when phrases are short. Second, generated utterances tend to end with a decreasing pitch, typical of the ending a sentence, rather than of an intermediate point. In the following section, we discuss the alternative approach of generating the entire sentence by forcing a pause in correspondence to the phrase breakpoints through some punctuation marks (two commas) in the dubbed sentence. While all punctuations induce pauses, phrase separating punctuations typically induce pauses $\geq$ 150 ms.

3.2. Post-Segmentation Approach

After inserting pause marks in the translation in correspondence to the phrase boundaries chosen by the prosodic alignment algorithm, the Mel spectrum of the entire sentence is generated at normal speaking rate. The inserted pauses are detected through a forced alignment procedure (explained in the next subsection), which are used to isolate each phrase. Each phrase's spectrum is then morphed along time to match its target duration, before being transformed into speech signal through the vocoder (VC). Finally, the resulting speech signals are concatenated together with the required silence signals. While this approach has to rely on additional computations and the accuracy of the pause detection step, it has the advantage of potentially generating more natural speech. By vocoding the entire sentence, we alleviate the degradation observed with short phrase's vocoding and the problem of generating utterances ending with a decreasing pitch respectively. The proposed solution is illustrated at the bottom of the Figure 3.

Here, we present a method that identifies phrase separating pauses using the alignment of attention matrix and a refinement step. The advantages of using our method are: (i) Eliminating dependency on external alignment tools like Montreal Forced Aligner (MFA) [20], (ii) Comparable or better performance than MFA, (iii) Removal of extra vocoder (VC) step required by external alignment tools, and (iv) Optimization/training free for new languages. In the subsequent section, we present our method termed as Search-and-Refine algorithm.

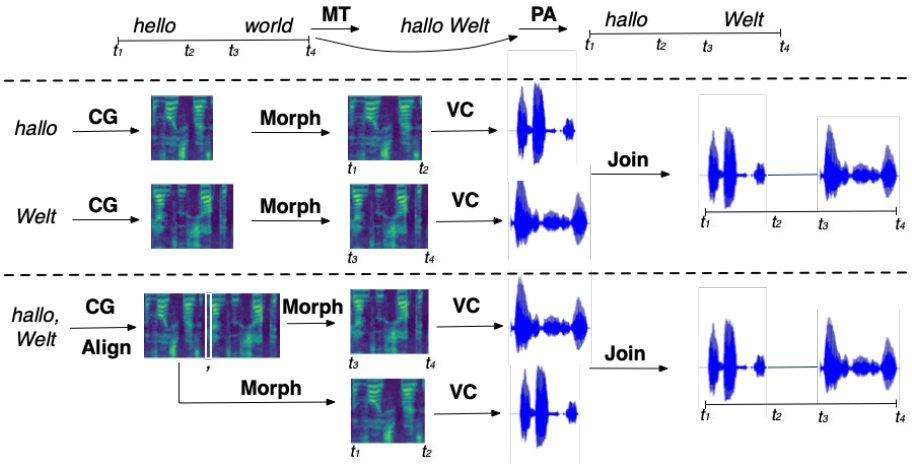

Figure 3: *Approaches to control TTS speaking rate at the phrase-level. (Top) Diagram showing how TTS input with timing information is generated by the prosodic alignment model. (Middle) Pre-segmentation approach: TTS and duration adjustment is applied on each input phrase. (Bottom) Post-segmentation approach: TTS is applied on the full input, after inserting a pause marker (comma), the output is segmented on the pause, and then duration is adjusted for each segment.*

3.3. Search-and-Refine Algorithm

To identify the time durations associated with the phrase separating pauses and other pauses, we follow a two step procedure.
1. Search step: We identify the most probable position of spectral frames corresponding to the inserted punctuation marks using Dynamic Time Warping (DTW) on the attention matrix. The attention matrix is observed to have similar probability values for several adjacent phonemes along the diagonal. This can potentially result in an incorrect alignment. We hypothesize that, for two adjacent phonemes having similar values, higher value should be assigned to the preceding phoneme. Following the correction, we use DTW to obtain the mapping corresponding to all the pauses. Further, if a punctuation precedes or succeeds a phoneme, we consider the maximum duration of that phoneme to be 2 frames [2] and assign the remaining duration to the frames associated with the punctuation. Following the procedure, we obtain start s_i and end frame e_i indices for the i^{th} the pause interval $[s_i, e_i]$. Subsequently, we use Mel spectrum energies to refine its output in the refinement step.
2. Refinement Step: For each start and end frame indices for the i^{th} pause interval $[s_i, e_i]$, we modify its duration by thresholding Mel spectrum energies in its neighbourhood of $k = 10$ frames.
The procedure consist of the following steps: First, we apply a median filter of length 3 to reduce noise in the Mel spectrum $S \in \Re^{L \times M}$ consisting of L frames and $M = 80$ bins. Second, we compute mean energy E_t for each frame t in the neighbourhood of the i^{th} pause interval as defined in the eq. 1:

$$E_t = \frac{1}{M} \sum_{j=1}^{M} S_t[j]^2, \; t = \{s_i - k, \ldots, e_i + k\}. \quad (1)$$

Third, we compute the energy derivative defined in the eq. 2:

$$D_t = \frac{\Delta E_t}{\Delta t}, \; t = \{s_i - k, \ldots, e_i + k\}. \quad (2)$$

[2] This duration corresponds to 125 ms, as the alignment map is downsampled by a factor of 5 with respect to the Mel spectrum with frame duration of 12.5 ms.

Fourth, we modify the pause or silence duration from search step to that of the local maxima/minima of the energy derivative. Finally, we merge two consecutive silence blocks if the difference between their end and start time respectively is smaller than 75 ms. We consider silence blocks with duration larger than 150 ms as phrase separating pauses. The amalgamation of *Search* and *Refinement* steps is known as Search-and-Refine algorithm.

4. Experiments

In this section, we present experimental results aiming at evaluating the accuracy and quality of the post-segmentation approach. In particular, we evaluate the accuracy of our segmentation procedure based on the Search-and-Refine algorithm over a multilingual benchmark. In our experiments, pauses in the target sentence were annotated using the prosodic alignment (PA) algorithm. We also compare the segmentation quality of our algorithm against MFA. Finally, we measure the quality of dubbed speech generated via pre- and post-segmentation over a multi-lingual benchmark in which a variety of TTS inputs for dubbing are simulated.

4.1. TTS Segmentation Quality

In this section, we describe the experiments to assess the quality of detecting pauses (both phrase separating and others). We compare the predicted timings of pause durations by our Search-and-Refine method and Montreal Forced Aligner (MFA) against reference pause durations. To obtain the reference pause durations, we use an energy-threshold method on the VC outputs using the audio processing tool Librosa [21]. We measure segmentation quality in terms of purity and coverage [22], being aware that 100% coverage is not necessarily in our reach, as the reference could include more silence frames than those we are targeting.

We assess the TTS segmentation quality for three languages (French, German and Spanish) using 50 sentences from our proprietary speech dubbing test set. For each of the 50 sentences in the three languages, we generate two sentences as follows: a) by inserting double commas at the phrase endings which were

Table 1: *Segmentation quality measured with percent coverage (C%) and percent purity (P%) of both pauses and speech on 100 sentences with 3 TTS voices for each language. MFA is the Montreal Forced Aligner output. * shows where differences with Search-and-Refine are statistically significant ($p< 10^{-4}$).*

Methods	French	German	Spanish
MFA (P%)	93.88	91.51*	93.75*
MFA (C%)	88.31	**92.42**	**94.19**
Search (P%)	80.62*	80.52*	92.33*
Search (C%)	83.84*	81.49*	90.25*
Search-and-Refine (P%)	92.60	**96.22**	**95.59**
Search-and-Refine (C%)	**88.39**	90.00	92.85

identified by Prosodic Alignment module and b) by inserting double commas at the phrase endings which were identified by manual/human segmentation. The generation step results in 100 sentences per language. We use the VC module to generate speech utterance with three voices per language (two female and one male voice) resulting in a total of 300 utterances per language. We obtain the reference pause durations using the energy-threshold method available in Librosa [21].

Table 1 presents the mean segmentation purity and coverage metrics for pauses and speech segments on 300 utterances, including 3 voices, for each language. We report the results of MFA, Search step and Search-and-Refine step described in the Section 3.3. We observe that the Search-and-Refine algorithm results in the highest purity for German and Spanish and highest coverage for French across the methods compared. We also observe that the refinement step always results in an improvement over the Search step. Enhancements over MFA are statistically significant ($p < 10^{-4}$) for Spanish and German, where significance tests were carried out with the one-sided Wilcoxon signed-rank test [23]. For French, our method results in metrics similar to those obtained by MFA. This empirical result demonstrates that external forced aligners are not required for the segmentation of the TTS output into pauses and speech segments.

As stated before, undetected pauses associated with phrase separating commas can result in the impossibility to properly segment the audio output. This phenomenon is known as under-segmentation. We observe that MFA misses 5.67%, 5.00% and 9.67% of the phase separating pauses in French, German and Spanish respectively. In contrast, our Search-and-Refine algorithm significantly improves over MFA in this respect by only missing, respectively, 0%, 0.33% and 0.67% of the pauses associated with the phrase separating commas. Nevertheless, such misses can always be detected and signalled to the downstream task so that countermeasures can be taken to properly segment the TTS output. For instance, possible fallback solutions are to use different punctuation symbols that can force a more prominent pause, such as semicolon or in the worst-case resort to the pre-segmentation approach. In the following section, we evaluate the subjective TTS output quality of the pre- and post- segmentation approach.

4.2. TTS Audio Quality

In this section, we compare the subjective audio quality of pre- and post-segmentation methods described in the Section 3. We selected 50 examples at random from our proprietary speech dubbing test set ensuring roughly 17 samples for every speaker in each of the three languages (French, German and Spanish).To simulate the dubbing scenario, for each phrase in each example, we select a speaking rate drawn randomly from the [0.9, 1.1] interval. Next, we generate the audio by morphing the corresponding speech spectrums obtained using both methods. In this experiment, we use a narrow band of speaking rates to prevent distracting the subjects from the task of evaluating audio quality. We asked 20 native speakers in each language to grade the naturalness of audio examples on a 0-10 scale. We collected a total of $1,000$ scores for each language and condition.

Table 2: *Subjective preferences for audio quality with using the pre and the post segmentation approaches. * marks improvements that are statistically significant ($p < 0.01$) .*

Methods	French	German	Spanish
Pre	23.86%	30.59%	24.73%
Post	38.26%*	38.59%*	29.47%

Results of subjective preferences are summarized in Table 2. The post-segmentation method is always preferred over the pre-segmentation method with relative differences of +60.4%, +26.2% and +19.2% for French, German and Spanish respectively. The improvements over pre-segmentation for French and German are statistically significant [24] ($p < 0.01$). A separate analysis for each voice and language also confirms that output from the post-segmentation method is preferred over pre-segmentation. The subjects additionally provided comments along the following key themes: (i) the post-segmentation method improves intonation, (ii) the voices sound robotic in a few cases, (iii) the changes in speaking rate and the placement of pauses were unnecessary in a few cases. We expect to address the second point by significantly improving the naturalness of our voices under different speaking rates. The last issue is likely due to the simulated dubbing scenario.

5. Conclusions

We discussed the problem of adjusting the speaking rate of neural TTS for phrases inside a sentence to meet the typical synchronization constraints posed by speech dubbing. We described the naive approach of generating and adjusting each phrase independently (pre-segmentation), and then concatenating them with the required pauses. Although simple to apply, this method might generate poor prosody at the end of phrases. Alternatively, we proposed generating the entire sentence after introducing punctuation at the end of each phrase and segmenting the spectrum by leveraging the attention matrix computed by the neural TTS model. Experiments carried out on three languages, with three voices each, show that Search-and-Refine can produce segmentations that are more comparable to those generated with the state-of-the-art Montreal Forced Aligner. Our method also results in a considerably lower percentage of missed segments. Finally, human evaluations show that speaking rate adjustments at the phrase-level obtained from post-segmentation approach result in more natural sentences than those generated with pre-segmentation.

6. References

[1] M. Federico, R. Enyedi, R. Barra-Chicote, R. Giri, U. Isik, A. Krishnaswamy, and H. Sawaf, "From Speech-to-Speech Translation to Automatic Dubbing," in *Proceedings of the 17th International Conference on Spoken Language Translation*, (Online), pp. 257–264, Association for Computational Linguistics, July 2020.

[2] F. Casacuberta, M. Federico, H. Ney, and E. Vidal, "Recent efforts in spoken language translation," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 80–88, 2008.

[3] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-Sequence Models Can Directly Translate Foreign Speech," in *Proc. Interspeech 2017*, pp. 2625–2629, ISCA, Aug. 2017.

[4] L. Cross Vila, C. Escolano, J. A. R. Fonollosa, and M. R. Costa-Jussà, "End-to-End Speech Translation with the Transformer," in *IberSPEECH 2018*, pp. 60–63, ISCA, Nov. 2018.

[5] M. Sperber and M. Paulik, "Speech Translation and the End-to-End Promise: Taking Stock of Where We Are," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 7409–7421, Association for Computational Linguistics, July 2020.

[6] F. Chaume, "Synchronization in dubbing: A translation approach," in *Topics in Audiovisual Translation* (P. Orero, ed.), pp. 35–52, John Benjamins B.V., 2004.

[7] A. Öktem, M. Farrùs, and A. Bonafonte, "Prosodic Phrase Alignment for Machine Dubbing," in *Proc. Interspeech*, 2019.

[8] M. Federico, Y. Virkar, R. Enyedi, and R. Barra-Chicote, "Evaluating and optimizing prosodic alignment for automatic dubbing," in *Proceedings of Interspeech*, p. 5, 2020.

[9] A. Saboo and T. Baumann, "Integration of Dubbing Constraints into Machine Translation," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, (Florence, Italy), pp. 94–101, Association for Computational Linguistics, Aug. 2019.

[10] S. M. Lakew, M. Di Gangi, and M. Federico, "Controlling the output length of neural machine translation," in *Proc. IWSLT*, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, pp. 5998–6008, 2017.

[12] M. Di Gangi, R. Enyedi, A. Brusadin, and M. Federico, "Robust neural machine translation for clean and noisy speech translation," in *Proc. IWSLT*, 2019.

[13] N. Prateek, M. Lajszczak, R. Barra-Chicote, T. Drugman, J. Lorenzo-Trueba, T. Merritt, S. Ronanki, and T. Wood, "In other news: A bi-style text-to-speech model for synthesizing newscaster voice with limited data," in *Proc. NAACL*, pp. 205–213, 2019.

[14] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, and K. Viacheslav, "Effect of data reduction on sequence-to-sequence neural TTS," in *Proc. ICASSP*, pp. 7075–7079, 2019.

[15] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, "Towards Achieving Robust Universal Neural Vocoding," in *Proc. Interspeech*, pp. 181–185, 2019.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. ICMAI*, pp. 234–241, Springer, 2015.

[17] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proc. ISMIR*, 2017.

[18] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. IWAENC*, pp. 1–4, 2010.

[19] E. A. Habets, "Room impulse response generator," Tech. Rep. 2.4, Technische Universiteit Eindhoven, 2006.

[20] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017* (F. Lacerda, ed.), pp. 498–502, ISCA, 2017.

[21] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," (Austin, Texas), pp. 18–24, 2015.

[22] H. Bredin, "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, (Stockholm, Sweden), August 2017.

[23] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*, pp. 196–202, Springer, 1992.

[24] E. W. Noreen, *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. John Wiley & Sons, 1989.