Joint phoneme alignment and text-informed speech separation on highly corrupted speech

Kilian Schulze-Forster, Clément Doire, Gael Richard, Roland Badeau

Accepted as conference paper at ICASSP 2020

JOINT PHONEME ALIGNMENT AND TEXT-INFORMED SPEECH SEPARATION ON HIGHLY CORRUPTED SPEECH

Kilian Schulze-Forster,[1*] *Clement S. J. Doire,*[2] *Gaël Richard,*[1] *Roland Badeau*[1]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] Audionamix, Paris, France

ABSTRACT

Speech separation quality can be improved by exploiting textual information. However, this usually requires text-to-speech alignment at phoneme level. Classical alignment methods are made for rather clean speech and do not work as well on corrupted speech. We propose to perform text-informed speech-music separation and phoneme alignment jointly using recurrent neural networks and the attention mechanism. We show that it leads to benefits for both tasks. In experiments, phoneme transcripts are used to improve the perceived quality of separated speech over a non-informed baseline. Moreover, our novel phoneme alignment method based on the attention mechanism achieves state-of-the-art alignment accuracy on clean and on heavily corrupted speech.

Index Terms— Speech separation, phoneme alignment, attention, informed source separation

1. INTRODUCTION

Speech separation research focuses mainly on noise as interfering source [1] which is highly relevant for applications such as telecommunication, hearing aids, or Automatic Speech Recognition (ASR) systems. Musical sound sources also often corrupt speech signals, which is relevant for separating speech in movies, radio shows, or home speaker speech recognition. The speech-music separation task has mainly been studied in simplified settings so far [2, 3].

State-of-the-art speech separation methods learn the task on large databases using deep learning [1]. A main challenge of purely data-driven approaches is the generalization to unseen data [4]. Therefore, it can be beneficial to exploit prior knowledge about the target speech source which is called informed source separation [5]. For example, a text transcript of the utterance in an observed mixture contains prior information about which sounds appear in which order in the speech source. It is often available for movies in the form of subtitles or scripts. Text transcripts have successfully been employed to improve speech separation [3, 6] but the methods have an important shortcoming: they rely on the availability of text that is aligned with the audio which is usually not the case. Thus, they perform automatic alignment as a pre-processing step, which comes with its own challenges. For example, methods for text-to-speech alignment are developed for rather clean speech and do not perform as well on corrupted speech as they are based on ASR [7, 8].

This results in a chicken and egg problem: alignment is required for text-informed speech separation and clean speech signals are required for high quality automatic alignment. Our hypothesis is that performing both tasks jointly leads to mutual benefits. The separation component facilitates alignment on corrupted speech while the alignment makes the text information more usable for the separation task. Apart from the separation task, aligning phonemes on corrupted speech has interesting applications such as generating training data for robust speech recognition systems or subtitles for movies.

We previously proposed a model for weakly informed source separation that can jointly exploit and align side information [9]. In this paper, we adapt the model for the text-informed speech-music separation task. We show that the perceived quality of the separated speech can be improved through text information without pre-alignment. Beyond, we propose a novel method for phoneme level text-to-audio alignment which achieves state-of-the-art performance on clean speech as well as on strongly corrupted speech with a Signal-to-Noise Ratio (SNR) of -5 dB. The alignment is derived from attention weights learned by the model in an unsupervised fashion when it is trained for the separation task.

2. RELATED WORK

Several deep learning approaches such as Long Short-Term Memory (LSTM) cells [4, 10], convolutional neural networks [11], and generative adversarial networks [12] have been applied to speech separation. LSTM networks have been shown to generalize well to unseen noises and speakers [4]. Text-informed speech separation has been studied first in [13]. An example speech signal is synthesized from the text and then aligned with the observed mixture using Dynamic Time Warping (DTW). The separation is done with a variant of non-negative matrix factorization exploiting similarities between the target speech and the example speech signal. The authors report that better alignment would have improved the results further. In [6], a sequence of phonemes is forced-aligned with noisy speech and then fed to a deep neural network together with the audio features. The authors show that the text information improves separation in terms of cepstral distance to clean speech. Information about phoneme identities is exploited for speech separation in [14] and [15] without using text-transcript as additional input. Instead, the phonemes are recognized using ASR techniques. Then, pre-trained phoneme-specific networks perform the separation. Additional effort is required to compensate for the limited performance of ASR on corrupted speech [14, 15].

Text-to-speech alignment faces two challenges: very long audio signals and corrupted speech. While some approaches cope with the former [7, 8], the latter is far from being solved for low SNRs. The method based on probabilistic kernels in [8] can align text with long audio signals but performance decreases when the speech is mixed with music. The approach in [16] applies ASR on a long speech signal and aligns a given text transcript with the recognized text. The

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowsa-Curie grant agreement No. 765068

process is iterated with an updated vocabulary and language model for regions that have not been aligned with high confidence in previous iterations. It can deal with noisy speech with an SNR of 15 dB. In [7], this approach is further improved by also updating the acoustic model on non-aligned regions leading to good alignment results up to an SNR of 10 dB.

The Montreal Forced Aligner (MFA) [17] goes even further. It uses a Gaussian Mixture Model (GMM) Hidden Markow Model ASR system and is trained in three iterative steps. First monophone, then triphone GMMs are trained iteratively, as in [7], to generate alignments on which acoustic feature transforms for speaker adaptation are learned as a third step.

The alignment capabilities of the attention mechanism have been already observed in [18] on a speech recognition task but have not been evaluated properly for alignment. Attention has been recently used to cope with non-aligned training data for a singing voice transcription task in [19].

3. METHOD

Let $x(t)$ be a linear mixture of speech and music with discrete time t. Let $Y = \{y_0, y_1, ..., y_{M-1}\}$ be a sequence of M phonemes represented as one-hot vectors which is a precise transcription of the utterance contained in $x(t)$. Our goal is to separate $x(t)$ into a clean speech signal $s(t)$ and background music $b(t)$ as well as predict onset times for each phoneme in the audio. Let $|X| \in \mathbb{R}^{F \times N}$ be the magnitude of the Short-Time Fourier Transform (STFT) of $x(t)$ with F frequency bands and N time frames. Given $|X|$ and Y, our model produces an estimate of the STFT magnitude for the clean speech source $|\hat{S}| \in \mathbb{R}^{F \times N}$. Considering that we will apply our model on low SNRs and given the results in [10], we directly output $|\hat{S}|$ instead of a mask. An inverse STFT of $|\hat{S}|$ combined with the mixture phase is performed to obtain the clean speech estimate in the time domain $\hat{s}(t)$.

As a starting point, we apply the model proposed in [9] for weakly informed audio source separation which is reviewed in Section 3.1. Adaptations for text-informed speech-music separation are detailed in Section 3.2. Further, we retrieve phoneme alignment information from the attention weights with the DTW algorithm as explained in section 3.3. A sketch of the workflow of our method is presented in Figure 1.

3.1. Weakly informed audio source separation

The model in [9] has two encoders which are two-layer bidirectional LSTM Recurrent Neural Networks (LSTM-RNN). The first encodes the side information sequence Y, the other encodes $|X|$, which can be seen as a sequence of N spectrogram time frames. The respective encoder outputs are the side information encoding $H = \{h_0, ..., h_{M-1}\}$ and the mixture encoding $G = \{g_0, ..., g_{N-1}\}$. An attention mechanism [20] is applied between H and G to find the relevant elements in the side information for each mixture encoding frame g_n. This allows the model to exploit side information without given alignment information. Specifically, a score is computed for all h_m with the learnable weight matrix W as

$$score_{n,m} = g_n^\top W h_m. \tag{1}$$

Attention weights $\alpha_{n,m}$ are then obtained by a softmax operation over the scores of all h_m:

$$\alpha_{n,m} = \frac{\exp(score_{n,m})}{\sum_{m=0}^{M-1} \exp(score_{n,m})}. \tag{2}$$

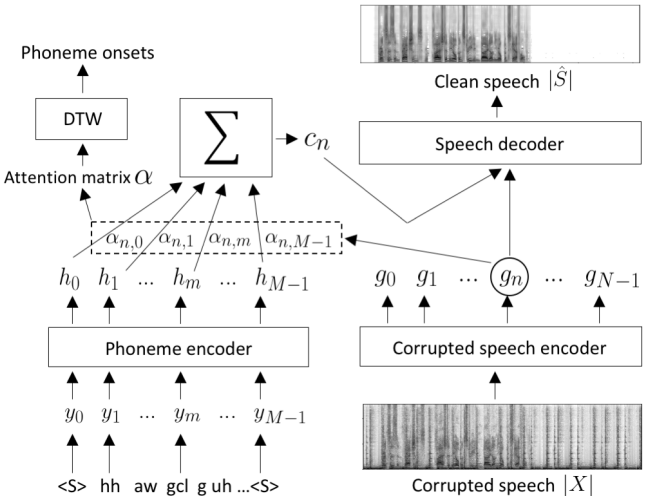

Fig. 1: Workflow of joint speech separation and phoneme alignment.

The attention weights reflect the relevance of the side information encoding element m for the mixture encoding at time step n. The side information is then summarized in a context vector c_n for each frame g_n individually as

$$c_n = \sum_{m=0}^{M-1} h_m \alpha_{n,m}. \tag{3}$$

The concatenation $[c_n, g_n]$ along the feature dimension is given to the target source decoder to compute the estimation $|\hat{S}|$. The decoder consists of a linear layer with $\tanh$ activation followed by a two-layer bidirectional LSTM-RNN and a linear layer with ReLU activation [9].

3.2. Adaption for text-informed speech-music separation

We derive three versions of the base model introduced above by modifying the way the phoneme sequence Y is processed, which we identified as a crucial point for the tasks at hand. It is worth mentioning that the phoneme encoding H serves two distinct purposes: (1) being an input to the attention mechanism identifying which phoneme is relevant at which mixture time frame and (2) being an input to the target source decoder in the form of c_n to inform the separation process.

For Version 1 (V1) we reduce the number of LSTM-RNN layers in the side information encoder to one and thereby limit its capacity. This leads to a more general representation of phonemes in H making it more applicable to fulfill its two purposes at once. Moreover, it reduces overfitting in limited data settings. Version 2 (V2) is identical to V1 except for an unidirectional LSTM-RNN in the phoneme encoder. This further reduces the number of learnable parameters and forces the model to process the phonemes strictly from left to right. Version 3 (V3) is equal to V1 but h_m is processed by a linear layer l before going into c_n. This changes equation (3) to

$$c_n = \sum_{m=0}^{M-1} l(h_m) \alpha_{n,m} \tag{4}$$

and means the model can learn two different representations of phonemes for their two purposes.

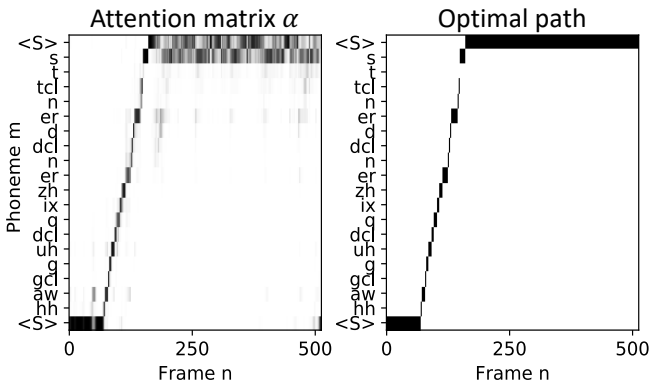

Fig. 2: Attention matrix (left) and DTW optimal path (right). Darker color represents higher values.

3.3. Retrieving phoneme onsets from attention weights

Given a sequence of phonemes Y and a corresponding audio signal $x(t)$ containing speech, the goal of phoneme-to-audio alignment is to estimate the onset times of each phoneme in the audio signal. We retrieve onsets from the attention weights using the DTW algorithm [21].

The attention weights can be represented collectively as attention matrix α with shape (M, N) as shown in Figure 2. With DTW, we find the optimal path through α from $(0, 0)$ to $(M-1, N-1)$ indicating which phoneme is active in which spectrogram frame. It maximizes the sum of attention weights it passes being restricted to only two possible moves, namely $(m, n+1)$ and $(m+1, n+1)$. This means we assume that all phonemes are pronounced and given in the correct order. An optimal path obtained this way is shown in Figure 2.

Knowing the hop length of the STFT that has been performed on $x(t)$ we know the exact position in time of each time frame n. The estimated onset time of a phoneme is the mid-point of the first time frame it has been assigned to by the optimal path.

Since our approach learns the alignment as a side outcome of learning speech separation, it can cope with much lower SNRs than other alignment methods which learn acoustic models from data directly. Training data with annotated phoneme onsets are not required.

4. EXPERIMENTS

We perform speech-music separation with joint phoneme alignment with the models V1, V2, V3 described in section 3.2. As baseline (BL) for the separation task, we use a model with the same configuration as V1. It resembles the speech separation model in [4] which is a four-layer LSTM-RNN with a linear output layer. Compared to [4], the BL has more expressive power through the attention mechanism and the phoneme encoder. It gets, instead of phonemes, a sequence of ones as side information, which does not convey any additional information about the speech signal to be separated. At the same time, BL has the same computational capacity as the models under test. This allows us to observe the exact effect of text as side information. We share the code of all models and experiments at `https://github.com/schufo/tisms`.

	SDR	SAR	SIR	STOI	PESQ
BL	8.81	**10.60**	14.53	0.87	2.66
V1	8.64	10.39	14.44	0.87	2.72
V2	**8.86**	10.57	**14.55**	**0.88**	**2.74**
V3	8.76	10.47	14.53	**0.88**	**2.74**
OA	*8.93*	*10.70*	*14.58*	*0.88*	*2.84*

Table 1: Separation quality evaluation results, all values are medians over the test set. SDR, SAR, SIR are shown in dB. BL: Baseline, V1-3: Version 1-3, OA: Optimal Attention weights.

4.1. Data set

We use the instrumental accompaniments of the MUSDB18 data set [22] as music signals and mix them with the TIMIT [23] speech signals. All music signals are converted to mono, downsampled to 16 kHz, and cut into snippets of 8.2 seconds, which is longer than all speech signals. For training, we mix snippets of 80 MUSDB tracks with 4320 utterances. The validation set contains 20 music tracks and 240 utterances and the test set 50 music tracks and 1344 utterances. The start time of the utterance within the 8.2 seconds of music is chosen randomly and differs for every example and every epoch. During training, we mix speech and music with a SNR uniformly drawn from [-8, 0] dB. For the validation and test set, we mix with SNR = -5 dB. The SNR is calculated only on the signal parts where the speech is active. There is no sentence or speaker overlap between the data sets. The STFT is computed with Fast Fourier Transform length 512, Hamming window, and hop length of 256 leading to magnitude spectrograms of size $(F \times N) = (257 \times 512)$. Each magnitude spectrogram is divided by its maximum value to normalize it to the range [0; 1].

As text information, we use the phoneme level transcripts that come with the TIMIT speech recordings. They comprise a vocabulary of 60 different phoneme symbols, to which we add a silence token (<S>) and a padding token for batching. The silence token is added to the start and end of each phoneme sequence because the speech is not active at the beginning and end of the mixture signal.

4.2. Training

We use the L1 loss, batch size 32, and the ADAM optimizer [24] with learning rate 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-6}$. The learning rate is reduced to 10^{-5} for the first 200 epochs. We stop training after 200 consecutive epochs without a decrease in validation cost.

5. RESULTS AND DISCUSSION

We will show that it is in fact beneficial to derive two phoneme representations as in V3 (cf. 3.2) if the model should produce good separation and alignment results at the same time. We will also show that performance on one task can be improved when neglecting the other task.

5.1. Speech-music separation results

We evaluate the predicted speech signals in terms of the BSS_eval metrics [25] Source-to-Distortion Ratio (SDR), Source-to-Artifacts Ratio (SAR), and Source-to-Interference Ratio (SIR) expressed in

	Clean speech		SNR = -5 dB	
	mean	median	mean	median
MFA	**16.3**	15.7	**38.4**	26.0
V1	22.5	**12.9**	39.0	**16.1**
V2	326.2	75.4	355.0	120.2
V3	48.1	14.4	69.0	17.9

Table 2: Mean Absolute Error (MAE) of phoneme onset predictions in ms averaged over the test set. MFA: Montreal Forced Aligner, V1-3: Version 1-3.

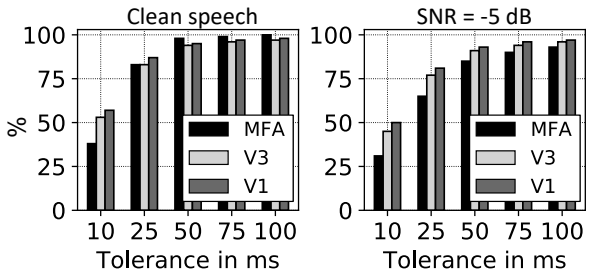

Fig. 3: Percentage of correctly aligned phonemes with different tolerances. MFA: Montreal Forced Aligner, V1-3: Version 1-3.

dB. We also compute the Perceptual Evaluation of Speech Quality (PESQ) [26] and the Short-Time Objective Intelligibility (STOI) measure [27]. SDR, SAR, SIR are computed on non-overlapping frames of 1 second length and the median value is taken to represent performance on one test example.

In Table 1, the median over the test set is presented for all metrics. Given the difficulty of the task (the SNR is -5 dB), BL performs well. The median STOI and PESQ of the corrupted speech in the test set are 0.64 and 1.48, respectively, which BL improves considerably. V1, V2 and V3 improve the PESQ over BL. This indicates that text information can improve the perceived quality of separated speech signals. V1 decreases the BSS_eval scores compared to BL, V2 changes them insignificantly, and V3 adds only slightly more artifacts than BL.

To test the upper bound of separation quality improvement through phoneme information in our experiment setting, we take our best model V2 and input the Optimal Attention (OA) weights during training and testing instead of learning them. We set $\alpha_{n,m}$ to 1 if phoneme m is active in frame n and to 0 otherwise based on true phoneme onsets. We can see in Table 1 that the SDR, SAR, and PESQ improve considerably over BL. This result shows that text information can improve speech separation even more when the alignment is provided.

Since the evaluation metrics do not capture all fine signal characteristics, we also provide some audio examples at `schufo.github.io/publications/2020-ICASSP`. In informal listening tests we observed that while some word endings are not audible in baseline predictions, they are clearly audible in the predictions of V2 and OA.

5.2. Phoneme-to-audio alignment results

As baseline for the phoneme alignment task, we use the MFA, which we reviewed in Section 2. To train it, we follow closely the procedure described in [17] which leads to advantageous conditions for the MFA: It is trained on the entire data set (including training, validation and test data), it gets the speaker identity of each utterance to perform speaker adaptation, and each example is cut at start and end of the utterance for training and testing (no long silent speech parts). We test all methods on the test set for two cases: clean speech and SNR = -5 dB. The MFA is trained on clean and corrupted speech respectively to learn appropriate acoustic models.

We evaluate the Mean Absolute Error (MAE) on each test example. It is the mean of the absolute differences between the true and predicted phoneme onsets in milliseconds (ms). The mean and median MAE over the test set are shown in Table 2. We see that V2 is not suited for phoneme alignment, whereas it performed best on the separation task. On clean speech, the MFA and V1 perform almost equally well. V1's median is better indicating that its alignments are more accurate when neglecting outliers. V1's mean is worse indicating that it produces more severe outliers. This can be explained by the dependence of our alignment method on somewhat sharp attention weights. When the model focuses on many phonemes at each time step n, i.e. α is not sharp, an optimal path indicating accurate phoneme onsets cannot be found.

On corrupted speech with SNR = -5 dB, V1 clearly outperforms the MFA. The mean of both methods is very similar while V1's median MAE is almost 10 ms lower. In general, V3 does not perform as well as V1 but still gives accurate predictions and outperforms the MFA regarding median MAE on clean and corrupted speech.

We also compute the percentage of correctly aligned phonemes within a tolerance around the true onsets. The results are shown in Figure 3. They confirm that V1's and V3's alignment accuracy is not much affected by strong speech corruption while the MFA's accuracy decreases. Moreover, V1 and V3 estimate more than 50 % of all phoneme onsets set with less than 10 ms error compared to the true onsets on clean speech. Even for the case of SNR = -5 dB, V1 aligns 50 % of the phonemes within the 10 ms tolerance.

6. CONCLUSION

Performing text-informed speech separation and phoneme alignment jointly leads to mutual benefits. After adapting a model for weakly informed source separation to perform both tasks, we showed that it can improve the perceived quality of separated speech with non-aligned phonemes as prior information. A novel phoneme alignment method based on attention arises from our joint approach. It achieves state-of-the-art accuracy on clean and on heavily corrupted speech.

7. REFERENCES

[1] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] Cemil Demir, Murat Saraclar, and Ali Taylan Cemgil, "Single-channel speech-music separation for robust asr with mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 725–736, 2012.

[3] Luc Le Magoarou, Alexey Ozerov, and Ngoc QK Duong, "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.

[4] Jitong Chen and DeLiang Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[5] Antoine Liutkus, Jean-Louis Durrieu, Laurent Daudet, and Gaël Richard, "An overview of informed audio source separation," in *14th International Workshop on Image Analysis for Multimedia Interactive Services*. IEEE, 2013, pp. 1–4.

[6] Keisuke Kinoshita, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani, "Text-informed speech enhancement with deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and S Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[8] German Bordel, Mikel Penagarikano, Luis Javier Rodríguez-Fuentes, Aitor Álvarez, and Amparo Varona, "Probabilistic kernels for improved text-to-speech alignment in long audio tracks," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 126–129, 2015.

[9] Kilian Schulze-Forster, Clément Doire, Gaël Richard, and Roland Badeau, "Weakly informed audio source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019.

[10] Lei Sun, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays*. IEEE, 2017, pp. 136–140.

[11] Se Rim Park and Jin Won Lee, "A fully convolutional neural network for speech enhancement," *Interspeech*, pp. 1993–1997, 2017.

[12] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "Segan: Speech enhancement generative adversarial network," *Interspeech*, pp. 3642–3646, 2017.

[13] Luc Le Magoarou, Alexey Ozerov, and Ngoc QK Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 2013, pp. 1–6.

[14] Zhong-Qiu Wang, Yan Zhao, and DeLiang Wang, "Phoneme-specific speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp. 146–150.

[15] Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger, "A phoneme-based pre-training approach for deep neural network with application to speech enhancement," in *IEEE International Workshop on Acoustic Signal Enhancement*. IEEE, 2016, pp. 1–5.

[16] Pedro J Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Fifth International Conference on Spoken Language Processing*, 1998.

[17] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, pp. 498–502.

[18] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[19] Ryo Nishikimi, Eita Nakamura, Satoru Fukayama, Masataka Goto, and Kazuyoshi Yoshii, "Automatic singing transcription based on encoder-decoder recurrent neural networks with a weakly-supervised attention mechanism," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 161–165.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[21] Taras K Vintsyuk, "Speech discrimination by dynamic programming," *Cybernetics and Systems Analysis*, vol. 4, no. 1, pp. 52–57, 1968.

[22] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[23] John S Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2001, vol. 2, pp. 749–752.

[27] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.