Phoneme Level Lyrics Alignment and Text-Informed Singing Voice Separation

Kilian Schulze-Forster, Clement S. J. Doire, Gaël Richard, Roland Badeau

***Abstract*—The goal of singing voice separation is to recover the vocals signal from music mixtures. State-of-the-art performance is achieved by deep neural networks trained in a supervised fashion. Since training data are scarce and music signals are extremely diverse, it remains challenging to achieve high separation quality across various recording and mixing conditions as well as music styles. In this paper, we investigate to which extent the separation can be improved when lyrics transcripts are used as additional information. To this end, we propose a joint approach to phoneme level lyrics alignment and text-informed singing voice separation. It is based on DTW-attention, a new monotonic attention mechanism including a differentiable approximation of dynamic time warping. Experimental results show that the method can align phonemes with mixed singing voice with high precision given accurate transcripts. It also achieves competitive results on challenging word level alignment test sets using less training data than state-of-the-art methods. Sequential alignment and informed separation lead to improved separation quality according to objective measures. Text information helps preserving spectral phoneme properties in the separated voice signals.**

***Index Terms*—Singing voice separation, lyrics alignment, monotonic attention mechanism**

I. Introduction

Singing voice separation is the task of isolating the vocals from the instrumental accompaniment in music recordings. It has user-oriented applications such as karaoke, remixing, up-mixing, and also serves as a pre-processing step for music information retrieval tasks such as singer identification or lyrics transcription. State-of-the-art performance is achieved by Deep Neural Networks (DNN) trained in a supervised way [1]–[3] which requires a dataset of music mixtures along with their corresponding isolated vocals stems. Obtaining such audio data is difficult due to copyright restrictions. The biggest publicly available dataset with somewhat realistic music mixtures is MUSDB [4] which consists of 150 (mainly western) rock-pop songs. Audio data for other music genres are even scarcer. Moreover, when the characteristics of mixtures at test time deviate from those used during training (e.g. quieter voice or noisy and reverberant live recordings) the performance of DNN-based methods decreases [5]. The question arises whether separation quality can be improved without access to more audio data. In this context, we investigate the use of lyrics transcripts to inform deep learning based singing voice separation. Words can be decomposed into phonemes, the smallest sound units of a language, which have distinct spectral characteristics [6]. They contain information about the sounds produced by a singer, e.g. voiced/unvoiced, phonetic class, and order of appearance.

We assume that phonemes need to be aligned with the observed mixture in order to inform the separation process. While great progress has been made regarding lyrics alignment at word level using resource intensive methods [7], [8], phoneme level alignment is rarely addressed although the methods in [7], [8] could be adapted to it. In fact, when phoneme alignment is required, they are often aligned manually [9], [10] or tools such as [11] are used [8], [12], [13] which employ acoustic models based on Gaussian Mixture Model - Hidden Markov Models (GMM-HMM) and do not work well on mixed singing voice as will be shown in Section V-B1.

Instead of adapting existing alignment methods, we introduce a new approach to lyrics alignment in this paper. The alignment is learned jointly with the informed separation. To this end, the source separation model Open Unmix [1] is extended so that it can process a phoneme sequence and a mixture as inputs. Driven solely by a separation objective, it learns to align text and audio with a new monotonic attention mechanism in order to derive a combined representation from which the voice spectrogram can be estimated. The idea is based on our previous findings [14], [15]. In [14] non-aligned voice activity information is exploited for singing voice separation and in [15] it is shown that phonemes can be aligned with corrupted speech signals via conventional attention [16] when a model is trained for text-informed speech-music separation [15]. However, the method in [15] does not work on singing voice mixtures without an important modification which we propose in this paper: the integration of Dynamic Time Warping (DTW) in the attention mechanism, which allows to exploit the left-to-right nature of text and audio sequences in order to obtain monotonic alignments. Competitive lyrics alignment performance is achieved with this joint approach although much less training data are used than for state-of-the-art methods [7], [8]. However, the separation performance does not improve compared to non-informed methods. Therefore, we also investigate a sequential approach where the phonemes are first aligned with our method and then fed as side information to a dedicated separation model. This leads to improvements through the text-information.

Manuscript received August 27, 2020; revised January 14, 2021 and May 3, 2021; accepted XXXXX XX, 2021. Date of publication XXXXX XX, 2020; date of current version XXXXX XX, 2020. 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. The associate editor coordinating the review of this manuscript and approving it for publication was XXXXX. *(Corresponding author: Kilian Schulze-Forster.)*

Kilian Schulze-Forster, Gaël Richard, and Roland Badeau are with Laboratoire de Traitement et Communication de l'Information (LTCI), Télécom Paris, Institut Polytechnique de Paris, France. (email: {kschulze, gael.richard, roland.badeau}@telecom-paris.fr)

Clement S. J. Doire was with Audionamix, Paris, France when the work was conducted and is now with Sonos Inc., Paris, France.

In summary, the contributions of this work are:

- a novel approach to lyrics alignment at phoneme level and lyrics-informed singing voice separation
- DTW-attention: a new monotonic attention mechanism including a differentiable approximation of DTW
- extension of the MUSDB dataset with lyrics transcripts and other annotations
- extensive experimental evaluation of the proposed method on lyrics alignment and informed singing voice separation

The paper is structured as follows. In Section II we review related work on lyrics alignment, monotonic attention, and text-informed source separation; the proposed method is explained in Section III. Data annotations and training details are presented in Section IV. The lyrics alignment evaluation is detailed in Section V, the singing voice separation evaluation is detailed in Section VI. The work is concluded in Section VII.

II. Related work

In this section, related work on lyrics alignment, monotonic attention, and informed source separation is reviewed.

A. Lyrics alignment

Most approaches to automatic lyrics alignment are based on acoustic models that estimate text unit (e.g. phoneme, character) probabilities given acoustic input features. When no large dataset of music recordings with corresponding lyrics was available, some acoustic models have been trained on speech and then adapted to singing voice [13], [17], but with limited success. Some works proposed to take additional information next to acoustic features into account such as chord labels [18] or phoneme durations inferred from a musical score [19]. These methods achieved good performance when aligning lyrics at phrase level on mixtures.

Recently, deep learning based approaches have exploited larger data resources for acoustic modeling on singing voice [7], [8]. They achieved high accuracy for word level alignment with mean absolute alignment errors below one second on mixed singing voice. The method of Stoller et al. [7] learns an acoustic model on time domain signals to estimate character probabilities over audio frames. It is trained on 39,232 songs with line level aligned lyrics using the Connectionist Temporal Classification (CTC) loss [20]. The data intensive nature can be explained by the end-to-end approach and the fact that character sequences are ambiguous regarding the word pronunciation.

Gupta et al. [8] proposed to learn three genre-specific acoustic models for the broad classes pop, hip hop, and metal on mixtures. Genre-specific models for non-vocal segments are learned to improve the performance on long instrumental parts. It requires a training corpus with genre labels and enough data per genre class to train all acoustic models. In total, 3913 songs are used for training. Acoustic modeling and alignment are done using the open source speech recognition toolkit Kaldi [21] with a duration-based pronunciation lexicon for singing voice [22]. The performance seems to rely on a very large beam width during Viterbi decoding [23] as mentioned in the previous work [24] which is computationally expensive.

Instead of adapting the data intensive methods [7], [8] for phoneme alignment, we propose a novel alignment approach. The proposed model is actually trained for informed source separation and learns the acoustic model without direct supervision as a side effect. It has the potential to reduce the amount of required training data compared to [7], [8] because the task it solves during training is simpler. It has to *match* the observed phoneme sequence with the observed audio frames, whereas the other models need to *classify* observed audio frames into phonemes. Multitrack data are required for training of the proposed method.

The Montreal Forced Aligner (MFA) [11] is a tool that can be used to learn GMM-HMM acoustic models and to align phonemes with audio signals. As initial alignment it assumes that all given phonemes belonging to a short audio example have the same length. On such an alignment a monophone GMM-HMM is trained while iteratively re-estimating the alignment. Then, triphone models are trained iteratively starting from the alignment provided by the monophone model. Speaker adaptation is performed as a last step if the speaker identities are known. The implementation is based on Kaldi [21]. Such a tool is commonly used to align phonemes with singing voice to prepare training data for other tasks [8], [12], [13]. Therefore, it will serve as one of the baselines for phoneme alignment.

B. Monotonic attention

The attention mechanism has been introduced by Bahdanau et al. [16] for neural machine translation with recurrent encoder-decoder models. It enables sequence-to-sequence models to evaluate the relevance of each element in one sequence with respect to the elements of another sequence by means of a learned scoring scheme. The scores can be viewed as alignment information for the two sequences. Attention has been shown to be useful in a wide range of tasks and model architectures [25]. In some cases, the alignment is known to be monotonic. Modifications to the attention mechanism have been proposed in the context of speech recognition [26]–[28] and machine translation [27]–[29] in order to enforce monotonic alignments which can help to disambiguate repeated elements in the sequences. We refer to such modified mechanisms as monotonic attention. One important difference between existing monotonic attention models and our model is that they consist of only one encoder and one decoder and the attention mechanism aligns the encoder output with hidden states of the decoder. The hidden states are computed autoregressively and cannot be observed all at once whereas we can observe both sequences to be aligned entirely because they are both inputs to the proposed model.

Chorowski et al. [26] proposed to consider the attention weights for the previous decoder time step in the scoring function for the current time step. This enables the model to learn a monotonic alignment but does not enforce monotonicity explicitly. Luong et al. [29] and Tjandra et al. [28] use a sliding window over the encoder output sequences and only compute attention weights for elements within this window. They explore both shifting the window monotonically from left to right over the encoder output and learning to predict the window position for each decoder step.

Raffel et al. [27] proposed a monotonic attention mechanism for online scenarios where the input to the encoder is observed step-by-step. They define a stochastic process modeling the dependency of the matching decision on previous time steps. It provides a hard alignment at test time and the model is trained using soft alignments which reflect the expected outcome of this process.

The sliding window approach and the stochastic process in [27] make the alignment decision at a certain time step dependent on decisions at previous steps. An incorrect matching at some time step can therefore lead to many incorrect matches at subsequent steps. Our approach relaxes the dependence of attention weights across time steps during training. At test time, DTW finds a *globally* optimal alignment which considers all elements of both sequences. Moreover, in autoregressive models the computation of attention weights cannot be parallelized for the decoder time steps. DTW-attention allows for more parallel computations.

Another difference to the typical single encoder-decoder attention mechanism is that in our model the information coming from the text is not essential (but potentially useful) in order to minimize the loss function, i.e. to learn the separation. Since the alignment is learned driven only by the separation objective, we observed that too strong constraints on the attention mechanism result in vanishing gradients for the text encoder and the attention mechanism so that no alignment is learned, while the separation is still learned. Therefore, the approaches proposed in [27] and [28] do not work in the context of this work. The proposed DTW-attention mechanism is able to learn the alignment while incorporating monotonicity constraints.

Cuturi et al. [30] proposed soft DTW which enables computing the DTW distance between two sequences with different lengths in a way that is differentiable and well-suited for gradient-based optimization. It allows using the DTW distance as a loss function but recovering the optimal *alignment path* is not possible. Therefore, soft DTW is not applicable in the context of this work and we propose DTW-attention to approximate the DTW *alignment path* in a differentiable way.

C. Informed audio source separation using deep learning

Prior knowledge has usually been used in source separation with Non-negative Matrix Factorization (NMF) in order to constrain the spectral templates and their temporal activations. The prior knowledge concerned either source characteristics (e.g. harmonic, continuous) or it was provided in the form of additional data related to the source signals, for example as musical scores [31].

DNNs have a larger capacity than NMF and achieve very good separation results without any extra information [1], [3]. However, recently there has been an interest in including side information such as pitch [32], [33] or phonetic content [34]–[37] in deep learning based separation in order to make it more robust in challenging scenarios. It has also been proposed to learn auxiliary tasks jointly, e.g. instrument activation detection [38] in order to cope with a larger number of musical sources to be separated. Most related to our work are four approaches that consider phonetic and linguistic information for singing voice separation.

Takahashi et al. [34] use deep features from an End-to-End Automatic Speech Recognition (E2EASR) model as side information for voice separation. The assumption is that the features contain phonetic *and* linguistic information because E2EASR combines acoustic and language modelling within one model. The side information leads to big improvements on speech separation in challenging conditions. The improvement for singing voice separation is considerably smaller. A possible reason is that the E2EASR model is trained on speech data and not adapted to singing voice.

Chandna et al. [35] train an encoder to extract content embeddings from mixtures. The target content embeddings are obtained with a speaker conversion method and contain phonetic information. From the embedding, a decoder estimates vocoder features which, along with a fundamental frequency estimate, are used to re-synthesise the voice signal from a mixture. The results show that the intelligibility of synthesized vocals is improved through phonetic features, but the overall subjective audio quality is lower than for filtering based separation methods.

An advantage of the approaches in [34], [35] is that no alignment method is required because phonetic information is extracted directly from the mixtures. On the other hand, the phonetic information is rather implicit and the mixture remains the only source of information. We consider explicit phoneme sequences from lyrics transcripts as additional model input that is independent from the mixture. Two approaches to lyrics-informed singing voice separation have been developed in parallel to our work.

Meseguer-Brocal et al. [36] use lyrics transcripts aligned at word level to condition singing voice separation with the U-Net [39]. Words are represented as bag of phonemes (without any temporal information at phoneme level) from which parameters are estimated to transform deep features in the U-Net encoder. Improvements over the classic U-Net are reported. However, it is not clear whether they are caused by the higher number of parameters in the conditioned U-Net, the voice activity information inherent in aligned text, or by the phonetic information. Since only word level alignment is available, the phonetic information of the text cannot be exploited entirely. Jeon et al. [37] condition singing voice separation on lyrics manually aligned at syllable level. They use a deep text encoder consisting of 1-D-convolutional highway layers [40]. The approach is evaluated on a private dataset of Korean amateur solo singing recordings mixed with unrelated accompaniments. To our understanding, only one singer sings at a time (no background singers, no multi-pitch singing). This facilitates learning the relation between phonemes and audio during training and the usage of text-information at test time. However, real commercial music recordings often contain multiple voices making the use of lyrics for separation less straightforward.

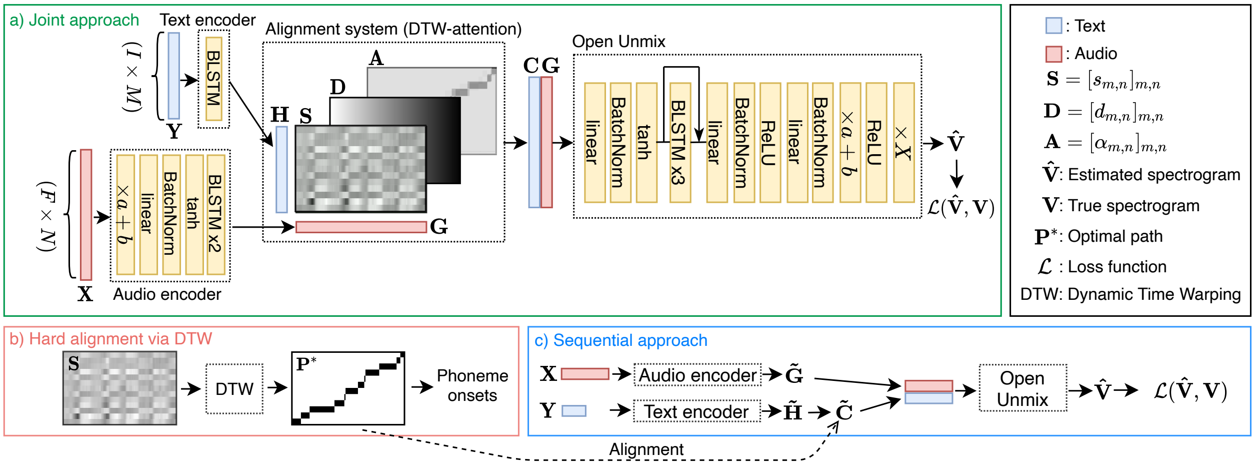

Fig. 1. Overview of the proposed model. **a)** With the joint approach, alignment and separation are learned by optimizing the separation objective. **b)** At lyrics alignment test time the phoneme onsets can be obtained from the score matrix via DTW. **c)** In the sequential approach, alignments are not learned but provided by some alignment method, e.g. the joint approach model.

In contrast to [36] and [37], we address the lyrics alignment problem which allows us to use lyrics aligned at *phoneme* level. Furthermore, we provide extensive experimental evaluation using publicly available realistic mixtures with multiple singers and correlated accompaniments. We conduct a thorough analysis of the separation performance regarding the number of simultaneously present singers and phonemes and regarding the Signal-to-Noise Ratio (SNR) of the voice-accompaniment mixtures.

III. Method

Notation: we denote scalars by italic lower and upper case characters (x, X), column vectors by boldface lower case characters $(\mathbf{x})$, and matrices by boldface upper case characters $(\mathbf{X})$. Matrix elements are denoted by scalars with two indices indexing the rows and columns respectively $(x_{a,b})$. A matrix may be treated as a sequence of column vectors. In this case, the vectors are indexed by the column number $(\mathbf{x}_n)$.

Let $x(t) = v(t) + a(t)$ be a time-domain single-channel mixture signal of singing voice $v(t)$ and instrumental accompaniment $a(t)$ where t refers to the discrete time index. Let $\mathbf{y} \in \{0, 1\}^I$ be a one-hot vector representing one out of I considered phonemes and let $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M] \in \{0, 1\}^{I \times M}$ be a matrix treated as a sequence of M one-hot vectors indexed by m representing the phonemes pronounced by the singing voice in the mixture.

The goal of text-informed singing voice separation is to separate $x(t)$ into $v(t)$ and $a(t)$ given $x(t)$ and $\mathbf{Y}$ as inputs. The goal of lyrics alignment is to estimate the onset time of each phoneme represented in $\mathbf{Y}$.

For this study, we propose a model that can perform lyrics alignment and singing voice separation jointly as well as sequentially. As shown in Figure 1, it consists of four parts: A text encoder and an audio encoder which are detailed in section III-A, an alignment system with a new monotonic attention mechanism explained in Section III-B, and Open Unmix [1] as a separation model described in Section III-C. A PyTorch implementation is available online[1].

A. The encoders

The text encoder is a single Bidirectional Long Short-Term Memory (BLSTM) layer [41], [42]. It transforms $\mathbf{Y}$ into the hidden phoneme representation $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_M] \in \mathbb{R}^{R \times M}$ where R is the number of hidden features.

In the audio encoder, the Short Time Fourier Transform (STFT) of the mixture signal $x(t)$ is computed and we denote its magnitude $\mathbf{X} \in \mathbb{R}_{\geq 0}^{F \times N}$ where F is the number of frequency components and N is the number of time frames which are indexed by $n = 1, ..., N$. Each time frequency bin is scaled and shifted by learnable scalars which are initialized by the standard deviation and mean over the training data, respectively, as in the Open Unmix model [1]. The audio encoder transforms the input with a fully connected layer with *tanh* activation followed by two BLSTM layers into the audio representation $\mathbf{G} = [\mathbf{g}_1, ..., \mathbf{g}_N] \in \mathbb{R}^{S \times N}$ where S is the number of hidden features.

B. The alignment system

The alignment system learns to align the vector sequences $\mathbf{H}$ and $\mathbf{G}$. An alignment can be formalized as a path denoted as sequence $\mathbf{P} = (\mathbf{p}_1, ..., \mathbf{p}_L)$ of length L with $\mathbf{p}_l = (m_l, n_l) \in [1 : M] \times [1 : N]$ that satisfies the following conditions [43]:

$$\mathbf{p}_1 = (1, 1) \quad \text{and} \quad \mathbf{p}_L = (M, N) \tag{1}$$

$$\mathbf{p}_{l+1} - \mathbf{p}_l \in \{(0, 1), (1, 1)\}. \tag{2}$$

Each path tuple $\mathbf{p}_l$ matches one phoneme with one audio frame. The step size condition in (2) is chosen so that each audio frame is matched with exactly one phoneme, whereas the same phoneme can be assigned to several audio frames. It follows that $L = N$. The condition also implies that the

[1] https://github.com/schufo/plla-tisvs

alignment path is monotonic and continuous, i.e. we assume that the phonemes are pronounced in the given order and no phoneme is skipped. The goal is to find the path that provides the correct matching between audio and text.

When phonemes are to be aligned with speech, a standard attention mechanism can learn such a monotonic alignment [15]. However, when working on singing voice, we found it to be crucial to enforce monotonicity explicitly in order to learn an alignment. This is probably due to the wider range of possible acoustic realisations of phonemes in singing due to a wider pitch range and artistic expressiveness. Therefore, we propose DTW-attention, a combination of DTW and attention to obtain monotonic alignments.

First, we compute a pair-wise matching score $s_{m,n}$ between all elements of the sequences $\mathbf{G}$ and $\mathbf{H}$ as

$$s_{m,n} = \mathbf{g}_n^\top \mathbf{W} \mathbf{h}_m \tag{3}$$

with the learned weight matrix $\mathbf{W} \in \mathbb{R}^{S \times R}$ as typically done in attention mechanisms [29]. It evaluates how likely it is that the m-th phoneme is pronounced in the n-th audio frame regardless of the position of $\mathbf{g}_n$ and $\mathbf{h}_m$ in their respective sequence.

Then, we incorporate the conditions (1) and (2) by computing the accumulated score matrix $\mathbf{D} = [d_{m,n}]_{m,n} \in \mathbb{R}^{M \times N}$ as typically done in DTW as follows [43], [44]:

$$d_{m,n} = s_{m,n} + \max(d_{m,n-1}, d_{m-1,n-1}) \tag{4}$$

with

$$d_{0,0} = b \quad \text{and} \quad d_{0,n} = d_{m,0} = -\infty \quad \forall m, n > 0 \tag{5}$$

where b is a sufficiently large number. Note that in (4) the objective is to maximize the accumulated *score*, whereas classical DTW usually minimizes a *distance* [44]. The reason for this is that stronger similarity between a phoneme and an audio frame results in a *higher* score $s_{m,n}$ while it would produce a *lower* distance value. The value $d_{m,n}$ is the accumulated score of the optimal alignment path starting at $(1,1)$ and ending in (m,n) respecting the step size condition (2). The optimal path in the DTW sense is the one with the highest accumulated score. The DTW step in (4) helps disambiguate identical phonemes appearing several times in the sequence, which could have the same score at a given time frame, by explicitly taking their order into account. It can be implemented efficiently by parallelizing computations of entries on the anti-diagonal of $\mathbf{D}$ or those lying on a line parallel to it because they are mutually independent.

Using classical DTW the actual optimal path could now be found by path backtracking [43]. However, such hard alignment, where one audio frame is matched with exactly one phoneme, is not differentiable [27], [30] and thus not applicable in a deep learning model during training. Instead, we will use a soft alignment strategy during training which we explain in III-B1. When phoneme onsets are to be retrieved at test time, we are able to compute $\mathbf{P}$ using the scores $s_{m,n}$ and classical DTW as detailed in III-B2.

1) Training: We compute attention weights α by a column-wise softmax operation on $\mathbf{D}$ as typically done in attention mechanisms [16]:

$$\alpha_{m,n} = \frac{e^{d_{m,n}}}{\sum_{k=1}^{M} e^{d_{k,n}}}. \tag{6}$$

The M attention weights corresponding to audio frame n can be interpreted as a probability distribution over all phonemes for this time frame and hence provide a soft alignment. The phoneme with the highest accumulated score in frame n has the highest probability α. This is a local approximation of the globally optimal path that would be obtained by DTW. It assumes that the phoneme with the highest accumulated score at frame n will be part of the optimal path. As we explain in Section VI-B1, this is true for 84% of the frames on our test set. Equations (4) and (6) put a soft constraint on the attention weights to be monotonic, i.e. respecting (2). It is *soft* because the dependence between time frames is reflected only in (4) whereas the attention weights are computed for each frame independently in (6). This is in contrast to other methods for monotonic attention, which we reviewed in Section II-B, and avoids error propagation from previous frames at the cost that there is no guarantee for strict monotonic paths during training. We found this trade-off to be appropriate in order for the model to learn the correspondence between phonemes and spectrogram frames of (mixed) singing voice. It also allows for efficient parallel computation of attention weights. The attention mechanism does not require training data with aligned phonemes. However, if such data were available they could be exploited trough a supervised loss term on the scores or attention weights.

The text information corresponding to an audio frame is then computed as

$$\mathbf{c}_n = \sum_{m=1}^{M} \mathbf{h}_m \alpha_{m,n} \tag{7}$$

and a new text sequence $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_N] \in \mathbb{R}^{R \times N}$ which has the same length N as the audio sequence $\mathbf{G}$ is obtained. Finally, $\mathbf{C}$ and $\mathbf{G}$ are concatenated along the feature dimension and this combined text and audio representation is then processed further by the separation model as explained in Section III-C.

2) Hard alignment at test time: Once the model is trained, the scores $s_{m,n}$ can be used as a similarity measure between a given phoneme sequence and the spectrogram frames. A globally optimal alignment $\mathbf{P}^*$ can then be found by DTW which consists of (4) and path backtracking [43]. The path $\mathbf{P}^*$ is a hard alignment as it assigns exactly one phoneme to each audio frame. While a hard alignment is required to infer phoneme onsets at test time, the soft alignment provided by (4) and (7) can be used to inform the separation model at test time in order to have the same behaviour as during training. The estimated phoneme onset is the start time of the first frame it has been assigned to. An example of the scores and a DTW path is shown in Figure 2.

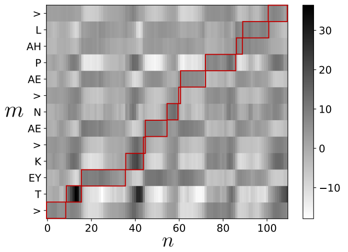

Fig. 2. Example of a score matrix $\mathbf{S} = [s_{m,n}]_{m,n}$ with optimal DTW path in red which assigns one phoneme to each audio frame.

C. The separation model

This part consists of the source separation model Open Unmix [1]. The input is the combined text and audio representation (cf. III-B1) from which the estimate $\hat{\mathbf{V}} \in \mathbb{R}^{F \times N}_{\geq 0}$ of the singing voice's magnitude spectrogram is computed. The model comprises a fully connected layer with $\tanh$ activation, three layers of BLSTM with a skip connection, and two fully connected layers with ReLU activation. The output is multiplied with the mixture magnitude spectrogram $\mathbf{X}$ and yields $\hat{\mathbf{V}}$. More details about the architecture are presented in Figure 1. In order to obtain the vocals estimate in the time domain, $\hat{\mathbf{V}}$ is combined with the mixture phase and an inverse STFT is applied. In this study, we do not consider additional models to estimate the other sources because the focus is on the effect of text-information for the vocals estimate.

By thresholding $\sum_{f=1}^{F} \hat{v}_{f,n}$ for each time frame n, the estimate can be used as a Voice Activity Detector (VAD) to find frames which are likely to contain no vocal sounds. At test time, the scores of space tokens that represent silence between words (cf. Section IV-B) can be set to a high value for such frames before applying DTW. This reduces the probability that phonemes are assigned to frames without vocals which can happen especially on long instrumental parts.

D. Joint vs. sequential approach

The model described above performs separation and alignment jointly. However, it can be beneficial for the separation quality to perform these tasks sequentially. For a sequential approach, two different, specialized versions of the model are employed. The first one (alignment model) corresponds exactly to the model described above. It is responsible for the alignment, which is learned through the separation objective as described. It is trained first and provides the hard alignment paths $\mathbf{P}^*$ for the second version (separation model) which is responsible for the separation and does not have an alignment system. We denote representations in the separation model with a tilde $\tilde{}$. The aligned text representation $\tilde{\mathbf{C}}$ is obtained by assigning an element of $\tilde{\mathbf{H}}$ to each audio frame using $\mathbf{P}^*$ (cf. Figure 1 and 2). During training and testing of the separation model, the text and audio sequences are fed to both the encoders of the alignment model and the encoders of the separation model (cf. Figure 1). The encoders of the separation model can learn representations $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{G}}$ dedicated exclusively to the separation task. In contrast, the representations in the alignment model and the model for a joint approach have to enable the alignment as well.

IV. Data annotation and training details

In order to obtain training and testing data for text-informed singing voice separation, we annotated the most popular singing voice separation dataset, MUSDB [4], with line level aligned lyrics and additional information about the vocals stems as explained in Section IV-A. We detail the training data and procedure in Section IV-B and a study on pre-training and attention is presented in Section IV-C.

A. Annotations of the MUSDB corpus

The dataset comprises 150 songs and is split into a training partition with 100 songs, of which 96 have English lyrics, and a test partition with 50 songs, of which 45 have English lyrics. We transcribed the English lyrics manually by listening to the vocals stems.

The songs were divided into sections of lengths between 3 and 12 seconds. The priority when choosing the section boundaries was that they correspond to natural pauses and do not cut vocal sounds. Most of the sections do not overlap, some have an overlap of 1 second. For each section, we annotated the start and end times, the corresponding lyrics as well as a label indicating one of the following four properties:

(a) only one person is singing
(b) several singers are pronouncing the same phonemes at the same time (possibly singing different notes)
(c) several singers are pronouncing different phonemes simultaneously (possibly singing different notes)
(d) no singing

Differentiating between singing voice examples with these properties allows for a more thorough analysis of the separation results and one could exclude certain segments from the training set, if desired. Segments that are labelled with the property (b) or (c) do not necessarily have this property over the whole segment duration. As soon as somewhere in a segment several singers are present, label (b) was assigned; as soon as they sung different phonemes somewhere at the same time, label (c) was assigned. Property (a) and (d) are valid for the entire segment. Furthermore, segments with property (c) can contain either some (lead) singer(s) singing some words in the presence of background singers singing long vowels such as 'ah' or 'oh' or they can contain multiple singers who sing different words at the same time. In the latter case, it was very difficult to understand the lyrics and to decide in which order to transcribe words or phrases sung simultaneously. We marked these segments and excluded them from our training and test data. In some difficult cases, e.g. shouting in metal songs or mumbled words, where the words are barely intelligible, we made an effort to make the transcriptions as accurate as possible phonetically and did not prioritize semantically meaningful phrases.

We believe that these annotations are a valuable resource for research on several tasks such as automatic lyrics alignment and transcription, text-informed singing voice separation, and singing voice analysis. Therefore, we make them publicly available[2].

B. Training details

We use 82 songs (2289 segments with total length of 4.6 hours) of the annotated MUSDB training set for training. The remaining 14 songs are used as a validation set (487 segments with 0.94 hours total length) for early stopping. The audio signals were downsampled to 16 kHz. As for the original Open Unmix model [1], training is done on short segments to prevent learning difficulties with backpropagation through time [45]. This does not prevent the model to process longer sequences at test time. Preliminary experiments (cf. Section IV-C) showed that the attention mechanism requires pre-training with mixtures containing speech signals. We found that pre-training on speech-music mixtures for 66 epochs enables subsequent training on singing voice plus accompaniment mixtures. We use speech recordings sampled at 16 kHz and word level text transcripts from the TIMIT database [46]. The speech is mixed with instrumental music retrieved from Youtube with a SNR uniformly drawn from $[-8,\ 0]$ dB. In total, the set consists of 4320 mixtures, which are between 2 and 8 seconds long and have a total length of 4.9 hours.

All words in the transcripts are translated into phonemes using the CMU LOGIOS Lexicon Tool[3]. Hence, there is no guarantee that the phonetic transcription always reflects the actual word pronunciation in the recordings accurately. We add a space token between each word that represents potential silence in the vocals. Examples without vocals are annotated with only the space token as lyrics.

The model is trained with the objective to minimize the L1 distance between the estimated and true vocals magnitude spectrogram, $\hat{\mathbf{V}}$ and $\mathbf{V}$ respectively. The ADAM optimizer [47], a learning rate of 0.001 and a batch size of 16 are used. A STFT with a Hann window of length 512 samples (32 ms) and a hop size of 256 samples (16 ms) is applied to compute the spectrograms. The learning rate is multiplied by 0.3 after 80 consecutive epochs without improvement of the validation loss and training is stopped after 140 consecutive epochs without improvement. Following the Open Unmix procedure, additive mixtures are produced for training by sampling the stems bass, drums, and others (as defined by MUSDB) randomly from different tracks, scaling them by a factor randomly drawn from $[0.25,\ 1.25]$ and adding them to a vocals segment scaled by a factor drawn from $[0.25,\ 0.9]$.

C. Study on pre-training and attention

In order to illustrate the effect of pre-training on speech-music mixtures, we train the proposed model with and without pre-training. To test the effectiveness of the proposed attention mechanism, we also train the model with a conventional attention mechanism [29] (applying the softmax operation in (6) on the scores $\mathbf{S} = [s_{m,n}]_{m,n}$ instead of the accumulated scores $\mathbf{D}$) for comparison. The resulting attention weights matrices for the four studied scenarios are shown in Figure 3.

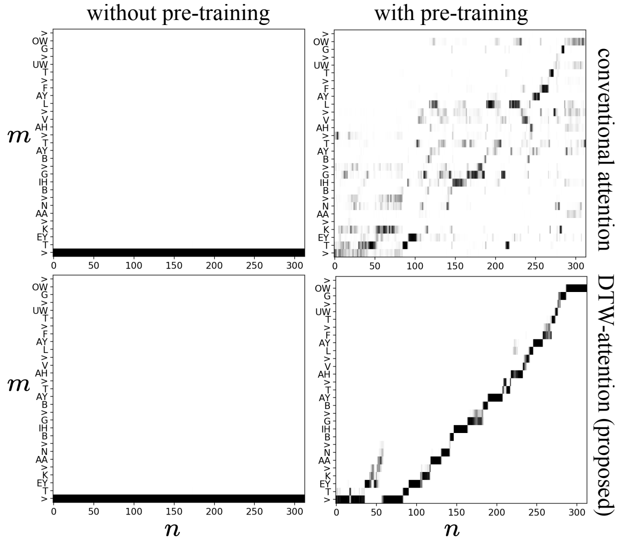

Fig. 3. Attention weight matrices $\mathbf{A} = [\alpha_{m,n}]_{m,n}$ for four different scenarios. Darker colors represent higher values, all values are in [0, 1].

Without pre-training on speech, none of the attention mechanisms learns an alignment for singing voice. With pre-training, both attention mechanisms learn some correspondence between audio and text, but only the proposed mechanism provides a sharp and nearly monotonic alignment. We can look at the differences between the speech and singing voice data used for training in order to understand why the attention mechanism initially requires speech data. The speech-music mixtures have more accurate text transcripts, a lower SNR (making the task more difficult and thus the side information more valuable), a lower voice pitch range, and more phonemes are uttered in a given time interval compared to singing voice. Also, word pronunciations are altered in singing voice for artistic reasons. We conducted various additional experiments with lower SNRs in the training examples using both the MUSDB data and singing voice recordings with accurate phoneme transcriptions [9]. In none of the settings did the attention mechanism train as desired. Therefore, the pitch range, phoneme rate, and uniform pronunciation in speech are likely to be the factors that enable the proper training in the considered limited data setting. A possible reason for the sensitivity to initialization is that the separation task can be learned by the model even without learning the alignment as discussed in the end of Section II-B.

The advantage of computing the attention weights for each audio frame independently from the other frames while still encouraging monotonicity can be seen in the bottom right plot of Figure 3: Although some phonemes are wrongly assigned to some early frames without singing, this mistake does not impede the correct monotonic alignment at later frames.

V. Alignment evaluation

We explain the experimental design for phoneme and word level lyrics alignment in Section V-A and present and discuss the results in Section V-B.

[2] https://doi.org/10.5281/zenodo.3989267

[3] http://www.speech.cs.cmu.edu/tools/lextool.html

A. Experimental design

Each test song is processed in full length at once by the model, so that no segmentation of audio and text is required, i.e. DTW is done on the score matrix $\mathbf{S} = [s_{m,n}]_{m,n}$ for the whole song.

1) Phoneme level alignment: We use the NUS-48E Sung and Spoken Lyrics Corpus [9] to assess phoneme level lyrics alignment. It is a collection of 48 solo singing recordings[4] of length between 53 seconds and 3.5 minutes with manually transcribed phonemes and their time stamps. 12 amateur singers sing 4 English songs each, the set comprises 20 unique songs. In order to evaluate phoneme alignment on mixtures, we mix each singing recording with an instrumental accompaniment of one song of the MUSDB test set.

We train the proposed model as explained in Section IV-B and call it JOINT1. Then, we test if some modifications regarding the training data can improve phoneme alignment. Since pre-training on speech data enabled learning the correspondence between phonemes and audio, speech data might also be beneficial when continuing training on singing voice. Therefore, we add 1000 speech-music mixtures to the MUSDB training data and call the model trained this way JOINT2. For training of the next model, we also add silence to the MUSDB vocals segments before mixing them with the other stems which results in longer instrumental sections in the training examples. This increases the amount of audio frames that correspond to the space token and potentially helps learning a better acoustic model for non-vocal frames. The idea is inspired by Gupta et al. [8] who identified acoustic modeling of non-vocal frames as a crucial aspect of automatic lyrics alignment. Specifically, each vocals signal is zero-padded to length 11 seconds. Padding is done for 50% of the signals at the start and for 50% at the end. The model trained with added silence and added speech is called JOINT3. We also train a model only on speech-music mixtures for comparison. It is called JOINT-SP.

Thereafter, we compare the best performing model from the study above to two baselines using both solo singing and mixtures as audio signals. The first one is the Montreal Forced Aligner (MFA) [11] (cf. II-A) which is a GMM-HMM. The MFA performs acoustic modeling and alignment iteratively and processes the training and test data combined. It is informed by the singer identity of the test songs and performs speaker adaptation. The second baseline is a deep learning model trained with the CTC loss [20]. It consists of three BLSTM layers with 256 hidden units followed by a linear layer mapping to the output size of 44 units (number of phonemes plus CTC's blank token). This architecture is inspired by the work in [48]. After a comprehensive hyper-parameter search, we found that the best performance on solo singing is obtained using 13 MFCCs (frame sise 256, 50% overlap) plus their deltas as input features. On mixtures it was best to use Mel-spectrograms (frame sise 512, 50% overlap) with deltas and delta-deltas as inputs. We call these versions CTC-MFCC and CTC-MEL, respectively. The model is trained with batch size 1 and a learning rate of 0.001. Both baselines are trained on our MUSDB training set. They are trained on mixtures for the evaluation on mixtures and on the clean vocals stems for the solo singing evaluation. Pre-training or including speech data or adding extra silence did not improve their performance.

2) Word level alignment: We evaluate word level lyrics alignment on the Hansen [49] and the Jamendo lyrics [7] dataset. They are widely used for word alignment evaluation on mixtures and comprise 10 and 20 western pop songs in English language, respectively. Also, they have been used in the Music Information Retrieval Evaluation eXchange (MIREX) 2019 lyrics alignment task[5], facilitating comparison between the proposed method and the two best performing methods which are from Gupta et al. (GU) [8] and Stoller et al. (ST) [7] reviewed in section II-A.

While word alignment can be considered as less difficult than phoneme alignment because it is coarser, these two datasets are more challenging than the one we have at our disposal for phoneme alignment. The reasons are that the accompaniment is correlated with the voice, they contain longer instrumental sections such as intros or solos, and the transcripts are partly incomplete as some vocal sounds such as 'ah' or 'oh' are sometimes neglected. Therefore, we also test using the vocals estimate $\hat{\mathbf{V}}$ as a Voice Activity Detector (VAD): when the estimated total vocal magnitude is lower than 20 for a time frame, it is assumed that it is a non-vocal frame and the score s of all space tokens is set to the maximum score obtained for the given song. This method is called JOINT3-VAD. The threshold was selected empirically on the MUSDB test set by visual inspection of the vocals magnitude for some examples. However, the alignment results have a marginal sensitivity regarding the exact threshold value because the VAD only reduces the largest errors as can be seen in Figure 4.

B. Results and discussion

1) Phoneme level alignment: The results of the experiment on training data are shown in Table I. The evaluation metrics are the mean and median Absolute Error (AE), which is the absolute difference between the true and estimated onset averaged over all phonemes of a test song, and the Percentage of Correctly Aligned Segments (PCAS) [50]. In this context, segments are the signal parts between onset time stamps and each segment is labelled with one phoneme. The PCAS measures the percentage of overlap of ground truth and estimated segments over the whole song. The AE compares onsets which are point estimates and does not take the phoneme duration into account whereas the PCAS tells which percentage of the audio signals is labelled with the correct phoneme. This is especially critical when the alignment is used for other downstream tasks such as informed separation in our case. Adding speech examples (+sp.) improves all evaluation metrics. There is less variance in the acoustic realisation of a phoneme in speech signals than in singing, which facilitates learning the relation between audio and phoneme

[4] We excluded song 09 of singer ADIZ due to incorrect annotations

[5] https://www.music-ir.org/mirex/wiki/2019:Automatic_Lyrics-to-Audio_Alignment_Results

TABLE I
PHONEME ALIGNMENT RESULTS ON NUS-48E CORPUS. VALUES ARE THE MEAN OVER THE TEST SET. AE=ABSOLUTE ERROR, PCAS=PERCENTAGE OF CORRECTLY ALIGNED SEGMENTS.

Method	Training data	mean AE [s]	median AE [s]	PCAS [%]	SNR [dB]
JOINT-SP	sp.	27.9382	26.5665	1.76	solo singing
JOINT1	MUSDB	0.0884	0.0158	81.49	
JOINT2	MUSDB+sp.	0.0611	**0.0149**	85.91	
JOINT3	MUSDB+sp.+sil.	**0.0573**	**0.0149**	**85.94**	
JOINT-SP	sp.	26.0748	23.7819	3.62	5
JOINT1	MUSDB	0.1122	0.0173	79.13	
JOINT2	MUSDB+sp.	0.0638	0.0160	84.41	
JOINT3	MUSDB+sp.+sil.	**0.0631**	**0.0158**	**84.66**	
JOINT-SP	sp.	33.4086	30.8661	0.97	-5
JOINT1	MUSDB	0.2639	0.0360	68.91	
JOINT2	MUSDB+sp.	0.1634	0.0254	75.38	
JOINT3	MUSDB+sp.+sil.	**0.1425**	**0.0247**	**76.02**	

TABLE II
PHONEME ALIGNMENT RESULTS ON NUS-48E CORPUS. VALUES ARE THE MEAN OVER THE TEST SET. AE=ABSOLUTE ERROR, PCAS=PERCENTAGE OF CORRECTLY ALIGNED SEGMENTS.

Method	mean AE [s]	median AE [s]	PCAS [%]	SNR [dB]
JOINT3	**0.057**	**0.015**	**85.94**	solo singing
MFA	0.073	0.030	77.94	
CTC-MFCC	0.071	0.034	76.49	
JOINT3	**0.063**	**0.016**	**84.66**	5
MFA	1.468	1.089	46.92	
CTC-MEL	0.198	0.078	57.61	
JOINT3	**0.077**	**0.018**	**82.17**	0
MFA	4.523	3.756	25.61	
CTC-MEL	0.513	0.267	46.94	
JOINT3	**0.143**	**0.025**	**76.21**	-5
MFA	7.079	6.172	10.03	
CTC-MEL	1.590	1.087	30.58	

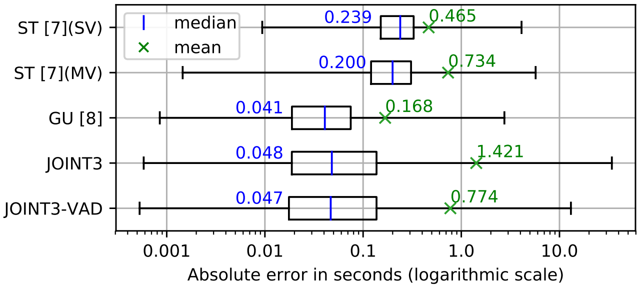

Fig. 4. Boxplot of the absolute alignment errors on the Jamendo data set [7]. The boxes extend from the first to the third quartile. The whiskers extend from the first to the 99th percentile.

labels statistically. Adding silence (+sil.) reduces the mean AE more than the median AE, and slightly improves the PCAS. As observed in [8], it helps recognizing non-vocal frames and makes the alignment more robust. Training only on speech-music mixtures (JOINT-SP) does not allow to align phonemes on singing voice. As a result of this study, we use the model JOINT3 for comparison with other methods on phoneme and word level alignment.

In Table II, JOINT3 is compared to the baselines MFA, CTC-MFCC, and CTC-MEL. The proposed method outperforms the baselines on solo and mixed singing voice. Note that the baselines have been trained on mixtures for the evaluation on mixtures (cf. V-A1). The fact that JOINT3 works well also on mixed singing, even with low SNRs shows the effectiveness of the voice separation component inherent in our alignment approach. In practice, the baselines could be used with voice separation as pre-processing step. However, it is likely that performance is worse than on solo singing. Comparing CTC-MFCC and JOINT3 shows that DTW-attention is more efficient in this limited data setting than CTC training. This can be explained by the fact that the CTC loss maximizes the likelihood of the target phoneme sequence given acoustic input features and marginalizes over all possible alignments. Therefore, the alignment that provides the correct frame/label synchronization is not preferred over other alignments. In contrast, the separation objective of the proposed method strongly favors the correct synchronization because it makes the phoneme information useful for the separation. The PCAS of the proposed approach is above 80 % for SNRs of 0dB and higher. This makes it a suitable method to produce phoneme alignments for datasets on which models for other tasks are trained.

2) *Word level alignment:* The word alignment results on the Hansen (H) and Jamendo (J) dataset are shown in Table III. The metrics are the mean and median Absolute Error (AE) (explained in V-B1) and the percentage of correctly aligned words within a tolerance of 0.3 seconds. A boxplot of the AEs on the Jamendo data set is shown in Figure 4. The mean and median values in the boxplot are taken over all AEs on the whole test set while the values in Table III are taken per song and are then averaged over all songs following the procedure of MIREX. Using the VAD reduces the mean AE and barely influences the median AE and overall error distribution. It can be seen in Figure 4 that the VAD decreases the largest errors. This happens because using VAD reduces the number of phonemes that are wrongly assigned to frames of long instrumental parts. We observed that changing the VAD threshold affects the results only marginally.

The baseline ST [7] has been evaluated for training and testing on Separated Vocals (SV) and Mixed Vocals (MV) by its authors. In Table III it can be seen that the baselines used considerably more training data than the proposed method. They have a lower mean AE than our method while the median AE is roughly the same. Figure 4 shows that the overall error distribution of the proposed method is very similar to the

TABLE III
WORD ALIGNMENT RESULTS ON THE HANSEN (H) [49] AND JAMENDO (J) [7] DATA SET. VALUES ARE THE MEAN OVER TEST SONGS.

Method	Songs for training	mean AE [s]		median AE [s]		% within 0.3s	
		H	J	H	J	H	J
ST [7] (SV)	39232	0.39	0.38	0.09	0.10	88	87
ST [7] (MV)	39232	-	0.82	-	0.10	-	85
GU [8]	3913	**0.10**	**0.22**	**0.04**	**0.05**	**97**	**94**
JOINT3	82[*]	1.47	1.86	0.06	0.10	83	80
JOINT3-VAD	82[*]	0.79	0.88	0.06	0.08	85	81

[*] plus 4.9 hours of speech music mixes (equals the length of 98 songs of 3 minutes)

state-of-the-art method GU [8] with the difference that some larger errors are produced which increase the mean AE. We observed that those outliers occur due to two reasons. Firstly, our method cannot cope well with vocal sounds that are not included in the given lyrics transcript because any vocal sound in an audio frame makes the model assign a higher score to the phonemes than to the space token (cf. equation (3)). This can result in assigning the first phoneme of the word after a non-transcribed sound to the frame of this non-transcribed sound and hence to predicting the onset too early. Secondly, the VAD does not capture all non-vocal segments perfectly and the model might confuse similar sounding instruments with vocals and assign high scores to phonemes instead of silence, which influences the DTW path. The baseline models learn more advanced acoustic models (triphone and genre-specific [8] or character level [7]) on more training data than our method. We think that this is the reason why they are more robust to those failure modes. They were the first to produce mean AEs below one second in the MIREX lyrics alignment task on mixtures. Considering the error distributions in Figure 4, the proposed method can be seen as a less data intensive alternative to the baselines. This is especially interesting for alignment of lyrics in other languages than English for which training data are scarcer.

To conclude the alignment evaluation, it can be said that the proposed method is able to align phonemes accurately on mixed singing voice when accurate transcripts are provided. Performance decreases when challenges such as long instrumental parts or inaccurate transcripts are faced, but performance is not far from the state-of-the-art on word level alignment in this case while less training data are used. DTW-attention trained with the separation objective yields better alignments than CTC training in the considered limit data setting.

VI. Singing voice separation evaluation

We explain the experimental design in Section VI-A and present and discuss the results in Section VI-B.

A. Experimental design

We use the 45 songs of the MUSDB [4] test set that are in English language along with their text transcripts for the separation evaluation. In total, our test set comprises 1461 segments with a total length of 2.9 h. The audio signals were downsampled to 16 kHz.

1) Open Unmix reference and joint approach: As a reference, we train the original Open Unmix model [1] on our MUSDB training data and call it UMX1. We also train it with the exact same training data and procedure as the best alignment model, JOINT3, i.e. pre-training on speech, adding silence to vocals, and adding speech data when training on singing voice (cf. Section V-A1), and call it UMX2. In order to evaluate the joint alignment and separation approach, we evaluate JOINT3.

2) Sequential approach: For the sequential approach (cf. section III-D and Figure 1), we use JOINT3 as the alignment model, providing alignments for a dedicated text-informed separation model which we call SEQ. Two baselines (BL) are provided. They use the exact same model as SEQ but, instead of one-hot vectors representing phonemes, they get different side information. For SEQ-BL1, every element in $\mathbf{Y}$ is the same one-hot vector and the given alignment path assigns the last element of $\mathbf{H}$ to all audio frames, i.e. $\mathbf{p}_n = (M, n) \forall n$. This means that no information about the singing voice is provided to SEQ-BL1. The second baseline, SEQ-BL2, receives the alignments provided by JOINT3 but all phonemes are represented with the same one-hot vector and the space token (silence) is represented with a different one-hot vector. This means the information of aligned phonemes is reduced to voice activity information for SEQ-BL2. Since the two baselines have the exact same architecture and number of parameters as SEQ, the effect of text as a side information can be evaluated.

3) Evaluation on mixtures with fixed SNR: For the experiments above, all models are evaluated with the original mixtures of the MUSDB dataset. Beyond that, we evaluate some models again and, this time, we mix the voice and accompaniment with a fixed SNR of 0, -5, and -10 dB. The SNR is computed on each test segment individually. This experiment allows us to investigate the effect of text as a side information on mixtures with different degrees of difficulty for singing voice separation. As reported in [5], lower SNRs usually decrease the separation quality.

B. Results and discussion

In Table IV, the separation evaluation scores are presented for several methods. The metrics are the Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), and the Source-to-Artefacts Ratio (SAR) [51] which are computed on one second long non-overlapping evaluation frames using *museval* with *BSSEval v4* [52] following the Signal Separation Evaluation Campaign [53]. We differentiate between the three annotated vocals properties of the test segments regarding the number of singers and the simultaneous presence of different phonemes (cf. Section IV-A). The values are the median over all evaluation frames within a property category. Higher values indicate better performance. Beyond, the Predicted Energy at Silence (PES) measures the energy of the estimated vocals in evaluation frames where the true vocals are all-zero, and the Energy at Predicted Silence (EPS) measures the energy in the true vocals for evaluation frames where the estimate is all-zero [14]. The presented values are the mean over *all* evaluation frames and lower values indicate better performance. The SDR, SAR, SIR are not defined for frames with a silent estimate or ground truth, so that the PES and EPS complement them for a complete evaluation. Note that a comparison of the presented performance scores with other models trained and tested on MUSDB is not straightforward because we were limited to the songs with English lyrics for training and testing.

1) Open Unmix reference and joint approach: The scores of UMX1 are lower than for the state-of-the-art version of Open Unmix [1]. The reason is the difference in training

TABLE IV
SEPARATION EVALUATION RESULTS IN DB. VALUES FOR SDR, SIR, SAR ARE MEDIANS OVER EVALUATION FRAMES, HIGHER VALUES ARE BETTER. VALUES FOR PES AND EPS ARE THE MEAN OVER EVALUATION FRAMES AND LOWER VALUES ARE BETTER.

			a) 1 singer			**b) 2+ singers 1 phon.**			**c) 2+ singers 2+ phon.**				
Method	Training data	Side info $\mathbf{Y}$	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	PES$\downarrow$	EPS$\downarrow$
UMX1	MUSDB	-	4.32	8.62	6.73	4.45	8.73	6.56	3.61	8.38	5.39	-72.26	-89.21
UMX2	MUSDB+sp.+sil.	-	4.06	8.62	6.22	4.31	8.30	6.56	3.85	7.87	5.69	-75.74	**-97.06**
JOINT3	MUSDB+sp.+sil.	phonemes	3.69	7.38	6.51	3.92	7.29	6.51	3.92	7.29	**6.17**	-84.09	-81.96
SEQ-BL1	MUSDB	constant	4.77	9.52	**7.16**	**4.93**	9.39	**6.91**	**4.20**	9.06	5.77	-93.57	-87.45
SEQ-BL2	MUSDB	voice activity	4.74	9.18	6.83	4.56	9.14	6.46	3.75	8.62	5.28	**-101.39**	-80.51
SEQ	MUSDB	aligned phonemes	**5.08**	**10.41**	6.82	4.89	**10.21**	6.51	3.86	**9.82**	5.03	-95.63	-85.98

($\downarrow$: lower values are better)

TABLE V
WORD ERROR RATE [%] OF THE LYRICS TRANSCRIPTION METHOD [54]

Method	Side info $\mathbf{Y}$	a)	b)	c)
Mixture		76.06	78.44	89.24
SEQ-BL1	constant	68.30	70.88	83.25
SEQ-BL2	voice activity	63.34	66.81	79.00
SEQ	aligned phonemes	**52.76**	**51.81**	**64.29**
True vocals		37.83	30.85	58.45

data such as the amount (we excluded non-English songs and multi-text segments), sampling rate, number of channels, and augmentation. In the original procedure, different random segments of 6 seconds length are cut out of the tracks at every epoch, whereas we are bound by the segment-wise aligned lyrics. However, this simulates the scenario which we investigate in this work: a limited amount of audio data available. Also, we focus on one model instance with vocals as target in order to investigate the effect of text as side information for the vocals estimate. In [1], four instances are used to estimate the four MUSDB targets which are combined using generalized Wiener filtering. UMX2 performs worse than UMX1. This indicates that the training data and process used for JOINT3 which enable the model to learn an alignment decrease the separation performance. The evaluation scores for JOINT3 show that the model has learned the separation task jointly with the alignment. However, the evaluation scores are lower than for the original Open Unmix model (UMX1 and UMX2). In the joint approach, the two encoders have to learn representations that enable both alignment and separation, which is worse for the separation than dedicated representations. JOINT3 was evaluated using the soft alignments provided by DTW-attention. However, using hard alignments of DTW instead has only marginal impact on the results. In fact, DTW-attention selects the same phoneme as the DTW path for 84% of all frames on the MUSDB test set, if we consider the phoneme with the highest weight as the one being selected, which is a reasonable assumption given the sharpness of the distribution (c.f. Figure 3). We conclude that joint alignment and separation is possible but not beneficial for the separation.

2) Sequential approach: The evaluation scores of the sequential approach SEQ are better than those for the joint approach, JOINT3. This is an expected result because dedicated representations can be learned by SEQ as discussed above. They are also better than those for the original Open Unmix model trained on the same data, UMX1. This improvement has two potential reasons: The proposed model has more capacity because of the two encoders and it uses text as additional information. We would like to know to which extent the performance increase is due to the text information. This can be seen when comparing SEQ to SEQ-BL1 and SEQ-BL2 which have all the same capacity. The text-informed model SEQ improves the SIR across all vocals properties compared to the less informed baselines. When only one singer is present (a) also the SDR is improved through the text information. The SAR is decreased when using text information with the decrease becoming stronger over categories a), b), and c). This shows that text information is most useful when only one person is singing. We discuss the limitations below in Section VI-B5. SEQ-BL2 has the lowest PES, which means it performs best on frames without vocals and thus uses the provided voice activity information. SEQ has the second lowest PES which indicates that it also uses the vocal activity information inherent in aligned text.

In order to evaluate an additional aspect of the separated singing voice signals, we automatically transcribe the lyrics from the voice estimates for the MUSDB test set using the state-of-the-art system [54] and compute the Word Error Rate (WER). The transcription system was trained on monophonic solo singing recordings of the Smule Sing! 300x30x2 dataset [55] and uses a language model built from lyrics [54]. The results in Table V show that the text-informed model SEQ produces a lower WER in each vocals category than the baselines. This means that the given phoneme information helps to preserve the characteristic phoneme properties in the separated voice signals.

An illustrative example is given in Figure 5. In the shown segment, a female singer sings the words ”right there almost got you”. The phonetic transcription of this line is ”> R AY T > DH EH R > AO L M OW S T > G AA T > Y UW >”, where ’>’ denotes the space token. The unvoiced ’s’ sound (in ”almost”) is missing in the estimate of the non-informed model (SEQ-BL1) but when using the text (SEQ) the model is able to separate it. Unvoiced sounds with high energy at high frequencies are difficult to differentiate from drum sounds such as cymbals, which makes text a valuable extra information. It can also be seen that the harmonic structure of the vowels is separated more clearly when using text. This leads to a clean sound and reduces interferences. However, it can also

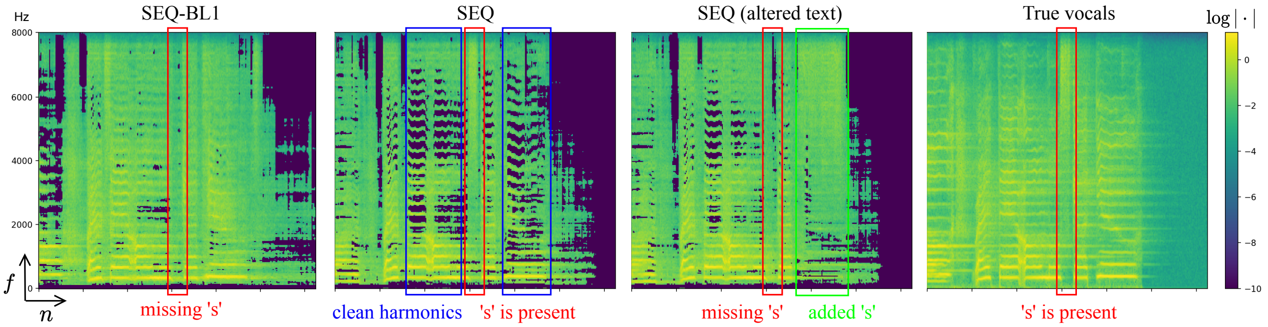

Fig. 5. Magnitude spectrograms of a singing voice obtained with different types of side information. SEQ-BL1: Meaningless side information, SEQ: aligned original phoneme sequence, SEQ (altered text): aligned modified phoneme sequence. On the right the true vocals are shown for comparison.

lead to artefacts, especially when multiple singers are present. Listening examples are provided online[6].

3) Relevance of the phonetic prior information: In order to test what kind of information is derived from the phoneme sequence by the model SEQ, we feed uniform white noise generated in the time-domain as input to the audio encoder at test time. We use phoneme sequences of the MUSDB test set as input to the text encoder. The alignment information of the phonemes with respect to their corresponding audio mixture is also provided. In the provided audio examples[6], it can be observed that the model filters the white noise so that the given phonemes become audible. The experiment shows that the model learned the spectral characteristics of the phonemes and how to use this information for the voice estimation. This explains why the separation with SEQ leads to a lower WER compared to the baselines.

Next, we test how much the model SEQ relies on the text information if it is conflicting with the observed audio mixture. To this end, we exchange some phonemes in the text after the alignment has been obtained with the original text. Using the example in Figure 5, we replaced the ’s’ in ’almost’ with an ’o’ so that its phonetic transcription became ”AO L M OW OW T” and we replaced the last word ’you’ by ”S S”. In Figure 5, it can be seen that the vocals estimate changes accordingly (SEQ (altered text)). The high frequency energy of the ’s’ sound is now missing where it was correctly estimated before (SEQ) for the word ’almost’. The spectral characteristics of ’s’ are added in the last frames where the word ’you’ was actually pronounced and where a clear harmonic structure was visible when the correct text was used (SEQ). We refer the reader to the audio examples for better illustration. This shows that the text information is actively used to estimate the voice and can even outweigh the information from the observed mixture. This can lead to a better separation as shown above but it can also lead to artefacts when the alignment or the transcription is inaccurate. Editing the phoneme sequence allows us to edit the obtained singing voice signal, e.g. to correct small pronunciation mistakes.

4) Evaluation on mixtures with fixed SNR: In Table VI, the separation results evaluated on mixtures with manually fixed SNRs are shown. The three annotated vocals properties of the test segments are differentiated and the values are the median over evaluation frames within each vocals category (a, b, c). For all SNRs and all vocals categories, the text-informed model SEQ achieves higher SIRs than both baselines. The SAR is reduced when using text information on all SNRs. The improvement of the separation through text becomes stronger when the SNR becomes lower. When the SNR is -10 dB also the SDR is clearly improved, even on test segments with multiple singers (b and c). We can conclude that text-informed singing voice separation is more beneficial in challenging conditions whereas it can lead to degraded performance in very easy conditions.

TABLE VI
SEPARATION EVALUATION RESULTS FOR MIXTURES WITH DIFFERENT SNRS. ALL VALUES ARE IN DB. EVALUATION SCORES ARE MEDIANS OVER EVALUATION FRAMES WITHIN A VOCAL CATEGORY.

		a)			b)			c)		
Method	SNR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
UMX1	0	8.00	12.65	10.60	7.16	11.87	9.75	4.90	10.35	6.93
SEQ-BL1		**8.59**	13.62	**11.04**	**7.57**	12.82	**10.00**	**5.43**	11.62	**7.43**
SEQ		8.36	**13.77**	10.27	7.34	**13.04**	9.51	4.94	**11.73**	6.43
UMX1	-5	4.45	8.47	6.93	4.08	8.12	6.18	2.51	6.15	4.17
SEQ-BL1		4.98	9.26	**7.32**	4.56	8.84	**6.55**	3.01	7.00	**4.93**
SEQ		**5.03**	**10.06**	6.83	**4.66**	**9.76**	6.29	**3.08**	**7.95**	4.10
UMX1	-10	0.91	4.01	3.46	0.81	3.39	2.42	0.03	0.65	1.78
SEQ-BL1		1.16	4.07	**3.91**	1.17	3.87	**3.09**	0.21	0.69	**2.42**
SEQ		**1.75**	**6.09**	3.38	**1.86**	**5.83**	2.87	**0.94**	**3.13**	1.74

5) Limitations: The discussion above in VI-B3 shows that accurate phoneme alignment and correct transcripts are necessary to achieve improvements through text. Otherwise, the vocals estimate will be degraded. In the case of multiple singers singing multiple phonemes (category c), the text contains information only about a part of the vocals signal, which is defined as mixture of all voice sources in MUSDB. Hence, the text-informed model SEQ as well as the model SEQ-BL2, which is informed by voice activity information derived from aligned text, might suppress the background singers when the lead vocals pause. Since the ground truth vocals contain all singers, this leads to lower evaluation scores. However, it can also be seen as an advantage if only the lead vocals are the source of interest. We refer the reader to the additional audio examples[7] to illustrate the points discussed above.

[6] https://schufo.github.io/plla_tisvs/

[7] https://schufo.github.io/plla_tisvs/

VII. Conclusion

The goal of this work was to investigate to which extent singing voice separation with deep neural networks can be improved through text information provided by lyrics transcripts. Since lyrics are usually not aligned with the observed mixture signals, we proposed a joint approach to phoneme level lyrics alignment and text-informed singing voice separation. Experimental evaluation showed that phoneme alignment can benefit from the separation component when the singing voice is mixed with other instruments. Moreover, the proposed alignment method achieved competitive results on two word level alignment test sets although it used less training data than state-of-the-art methods. In order to improve the separation performance, lyrics should be aligned first and subsequently be processed by a separation model. With this sequential approach, a text-informed model achieves higher separation quality than the baselines mainly in terms of SIR. The proposed model uses phoneme side information actively to shape the voice estimates. This preserves the phonetic information in the estimates but can also lead to degraded performance in case of inaccurate alignments or transcriptions. The impact of text is especially noticeable in challenging conditions such as low SNRs. One possible direction for future work is a comparison between phonemes and other types of side information such as pitch.

Acknowledgment

The authors would like to thank Emir Demirel for processing the data for the word error rate evaluation. We also thank Olumide Okubadejo and Sinead Namur for their help with transcribing the MUSDB lyrics.

References

[1] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, “Open-unmix - a reference implementation for music source separation,” *Journal of Open Source Software*, 2019. [Online]. Available: https://doi.org/10.21105/joss.01667

[2] N. Takahashi, N. Goswami, and Y. Mitsufuji, “MMdenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation,” in *Proceedings of the IEEE International Workshop on Acoustic Signal Enhancement*, 2018, pp. 106–110.

[3] A. Défossez, N. Usunier, L. Bottou, and F. Bach, “Demucs: Deep extractor for music sources with extra unlabeled data remixed,” *arXiv preprint arXiv:1909.01174*, 2019.

[4] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “The MUSDB18 corpus for music separation,” Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[5] K. W. E. Lin and M. Goto, “Zero-mean convolutional network with data augmentation for sound level invariant singing voice separation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 251–255.

[6] L. Rabiner and R. Schafer, *Theory and applications of digital speech processing*. Prentice Hall Press, 2010.

[7] D. Stoller, S. Durand, and S. Ewert, “End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 181–185.

[8] C. Gupta, E. Yılmaz, and H. Li, “Automatic lyrics transcription in polyphonic music: Does background music help?” *arXiv preprint arXiv:1909.10200*, 2019.

[9] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, “The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–9.

[10] R. Gong and X. Serra, “Singing voice phoneme segmentation by hierarchically inferring syllable and phoneme onset positions,” in *Proceedings of Interspeech*, 2018, pp. 716–720.

[11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi.” in *Proceedings of Interspeech*, 2017, pp. 498–502.

[12] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer,” in *Proceedings of Interspeech*, 2017.

[13] A. M. Kruspe, “Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing.” in *Proceedings of the International Society for Music Information Retrieval Conference*, 2016, pp. 358–364.

[14] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau, “Weakly informed audio source separation,” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019, pp. 273–277.

[15] ——, “Joint phoneme alignment and text-informed speech separation on highly corrupted speech,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[16] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proceedings of the International Conference on Learning Representations*, 2015.

[17] A. Mesaros and T. Virtanen, “Adaptation of a speech recognizer for singing voice,” in *Proceedings of the European Signal Processing Conference*, 2009, pp. 1779–1783.

[18] M. Mauch, H. Fujihara, and M. Goto, “Integrating additional chord information into HMM-based lyrics-to-audio alignment,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 200–210, 2011.

[19] G. B. Dzhambazov and X. Serra, “Modeling of phoneme durations for alignment between polyphonic audio and lyrics,” in *12th Sound and Music Computing Conference*, 2015.

[20] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi speech recognition toolkit,” in *Workshop on Automatic Speech Recognition and Understanding*, 2011.

[22] C. Gupta, H. Li, and Y. Wang, “Automatic pronunciation evaluation of singing.” in *Proceedings of Interspeech*, 2018, pp. 1507–1511.

[23] C. Gupta, E. Yılmaz, and H. Li, “AutoLyrixAlign: Pre-trained model and script to automatically align lyrics to polyphonic audio,” https://github.com/chitralekha18/AutoLyrixAlign, accessed: 2020-12-22.

[24] B. Sharma, C. Gupta, H. Li, and Y. Wang, “Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 396–400.

[25] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, “An attentive survey of attention models,” *arXiv preprint arXiv:1904.02874*, 2019.

[26] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[27] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, “Online and linear-time attention by enforcing monotonic alignments,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 2837–2846.

[28] A. Tjandra, S. Sakti, and S. Nakamura, “Local monotonic attention mechanism for end-to-end speech and language processing,” in *Proceedings of the International Joint Conference on Natural Language Processing*, 2017, pp. 431–440.

[29] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[30] M. Cuturi and M. Blondel, “Soft-DTW: a differentiable loss function for time-series,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 894–903.

[31] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, “An overview of informed audio source separation,” in *Proceedings of the IEEE International Workshop on Image Analysis for Multimedia Interactive Services*, 2013, pp. 1–4.

[32] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gomez, “Deep learning based source separation applied to choir ensembles,” in *Proceedings of the International Society for Music Information Retrieval Conference*, 2020.

[33] A. Jansson, R. M. Bittner, S. Ewert, and T. Weyde, “Joint singing voice separation and f0 estimation with deep u-net architectures,” in *Proceedings of the European Signal Processing Conference*, 2019, pp. 1–5.

[34] N. Takahashi, M. K. Singh, S. Basak, P. Sudarsanam, S. Ganapathy, and Y. Mitsufuji, “Improving voice separation by incorporating end-to-end speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 41–45.

[35] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, “Content based singing voice extraction from a musical mixture,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 781–785.

[36] G. Meseguer-Brocal and G. Peeters, “Content based singing voice source separation via strong conditioning using aligned phonemes,” *arXiv preprint arXiv:2008.02070*, 2020.

[37] C.-B. Jeon, H.-S. Choi, and K. Lee, “Exploring aligned lyrics-informed singing voice separation,” *arXiv preprint arXiv:2008.04482*, 2020.

[38] Y.-N. Hung and A. Lerch, “Multitask learning for instrument activation aware music source separation,” in *Proceedings of the International Society for Music Information Retrieval Conference*, 2020.

[39] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, “Singing voice separation with deep u-net convolutional networks,” in *Proceedings of the International Society for Music Information Retrieval Conference*, 2017, pp. 23–27.

[40] R. K. Srivastava, K. Greff, and J. Schmidhuber, “Training very deep networks,” in *Advances in Neural Information Processing Systems*, 2015, pp. 2377–2385.

[41] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[42] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with LSTM,” *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[43] M. Müller, *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

[44] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[45] P. J. Werbos, “Backpropagation through time: what it does and how to do it,” *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[46] J. S. Garofolo, “TIMIT acoustic phonetic continuous speech corpus,” *Linguistic Data Consortium, 1993*, 1993.

[47] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[48] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. D’alché-Buc, “Multilingual lyrics-to-audio alignment,” in *Proceedings of the International Society for Music Information Retrieval Conference*, 2020.

[49] J. K. Hansen, “Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients,” in *Proceedings of the Sound and Music Computing Conference*, 2012, pp. 494–499.

[50] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, “LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1252–1261, 2011.

[51] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[52] F.-R. Stöter and A. Liutkus, “museval 0.3.0,” Aug. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3376621

[53] F.-R. Stöter, A. Liutkus, and N. Ito, “The 2018 signal separation evaluation campaign,” in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 293–305.

[54] E. Demirel, S. Ahlbäck, and S. Dixon, “Automatic lyrics transcription using dilated convolutional neural networks with self-attention,” in *International Joint Conference on Neural Networks*, 2020, pp. 1–8.

[55] “Smule sing! 300x30x2 dataset,” https://ccrma.stanford.edu/damp/, accessed: 2020-01-11.