HAL
open science

CONTENT BASED SINGING VOICE SOURCE SEPARATION VIA STRONG CONDITIONING USING ALIGNED PHONEMES

Gabriel Meseguer-Brocal, Geoffroy Peeters

To cite this version:

Gabriel Meseguer-Brocal, Geoffroy Peeters. CONTENT BASED SINGING VOICE SOURCE SEPARATION VIA STRONG CONDITIONING USING ALIGNED PHONEMES. 21st International Society for Music Information Retrieval Conference, Oct 2020, Montréal (virtual), Canada. hal-03200161

HAL Id: hal-03200161

https://hal.science/hal-03200161v1

Submitted on 16 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CONTENT BASED SINGING VOICE SOURCE SEPARATION VIA STRONG CONDITIONING USING ALIGNED PHONEMES

Gabriel Meseguer-Brocal
STMS UMR9912, Ircam/CNRS/SU, Paris
gabriel.meseguerbrocal@ircam.fr

Geoffroy Peeters
LTCI, Institut Polytechnique de Paris
geoffroy.peeters@telecom-paris.fr

arXiv:2008.02070v1 [eess.AS] 5 Aug 2020

ABSTRACT

Informed source separation has recently gained renewed interest with the introduction of neural networks and the availability of large multitrack datasets containing both the mixture and the separated sources. These approaches use prior information about the target source to improve separation. Historically, Music Information Retrieval researchers have focused primarily on score-informed source separation, but more recent approaches explore lyrics-informed source separation. However, because of the lack of multitrack datasets with time-aligned lyrics, models use weak conditioning with non-aligned lyrics. In this paper, we present a multimodal multitrack dataset with lyrics aligned in time at the word level with phonetic information as well as explore strong conditioning using the aligned phonemes. Our model follows a U-Net architecture and takes as input both the magnitude spectrogram of a musical mixture and a matrix with aligned phonetic information. The phoneme matrix is embedded to obtain the parameters that control Feature-wise Linear Modulation (FiLM) layers. These layers condition the U-Net feature maps to adapt the separation process to the presence of different phonemes via affine transformations. We show that phoneme conditioning can be successfully applied to improve singing voice source separation.

1. INTRODUCTION

Music source separation aims to isolate the different instruments that appear in an audio mixture (a mixed music track), reversing the mixing process. Informed-source separation uses prior information about the target source to improve separation. Researchers have shown that deep neural architectures can be effectively adapted to this paradigm [1, 2]. Music source separation is a particularly challenging task. Instruments are usually correlated in time and frequency with many different harmonic instruments overlapping at several dynamics variations. Without additional knowledge about the sources the separation is often infeasible. To address this issue, Music Information Retrieval (MIR) researchers have integrated into the source separation process prior knowledge about the different instruments presented in a mixture, or musical scores that indicate where sounds appear. This prior knowledge improves the performances [2–4]. Recently, conditioning learning has shown that neural networks architectures can be effectively controlled for performing different music source isolation tasks [5–10]

Various multimodal context information can be used. Although MIR researchers have historically focused on score-informed source separation to guide the separation process, lyrics-informed source separation has become an increasingly popular research area [10, 11]. Singing voice is one of the most important elements in a musical piece [12]. Singing voice tasks (e.g. lyric or note transcription) are particularly challenging given its variety of timbre and expressive versatility. Fortunately, recent data-driven machine learning techniques have boosted the quality and inspired many recent discoveries [13, 14]. Singing voice works as a musical instrument and at the same time conveys a semantic meaning through the use of language [14]. The relationship between sound and meaning is defined by a finite phonetic and semantic representations [15, 16]. Singing in popular music usually has a specific sound based on phonemes, which distinguishes it from the other musical instruments. This motivates researchers to use prior knowledge such as a text transcript of the utterance or linguistic features to improve the singing voice source sparation [10, 11]. However, the lack of multitrack datasets with time-aligned lyrics has limited them to develop their ideas and only weak conditioning scenarios have been studied, i.e. using the context information without explicitly informing where it occurs in the signal. Time-aligned lyrics provide abstract and high-level information about the phonetic characteristics of the singing signal. This prior knowledge can facilitate the separation and be beneficial to the final isolation.

Looking for combining the power of data-driven models with the adaptability of informed approaches, we propose a multitrack dataset with time-aligned lyrics. We explore then how we can use strong conditioning where the content information about the lyrics is available frame-wise to improve vocal sources separation. We investigate strong and weak conditioning using the aligned phonemes via Feature-wise Linear Modulation (FiLM) layer [17] in U-Net based architecture [18]. We show that phoneme conditioning can be successfully applied to improve standard

© Gabriel Meseguer-Brocal, Geoffroy Peeters. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Gabriel Meseguer-Brocal, Geoffroy Peeters, "Content based singing voice source separation via strong conditioning using aligned phonemes", in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.,* Montréal, Canada, 2020.

singing voice source separation and that simplest strong conditioning outperforms any other scenario.

2. RELATED WORK

Informed source separation use context information about the sources to improve the separation quality, introducing in models additional flexibility to adapt to observed signals. Researchers have explored different approaches for integrating different prior knowledge in the separation [19]. Most of the recent data-driven music source separation methods use weak conditioning with prior knowledge about the different instruments presented in a mixture [3, 5, 7–9]. Strong conditioning has been primarily used in score-informed source separation. In this section, we review works related to this topic as well as novel approaches that explore lyrics-informed source separation.

2.1 Score-informed music source separation

Scores provide prior knowledge for source separation in various ways. For each instrument (source), it defines which notes are played at which time, which can be linked to audio frames. This information can be used to guide the estimation of the harmonics of the sound source at each frame [2,4]. Pioneer approaches rely on non-negative matrix factorization (NMF) [20–23]. These methods assume that the audio is synchronized with the score and use different alignment techniques to achieve this. Nevertheless, alignment methods introduce errors. Local misalignments influence the quality of the separation [21, 24]. This is compensated by allowing a tolerance window around note onsets and offsets [20, 23] or with context-specific methods to refine the alignment [25]. Current approaches use deep neural network architectures and filtering spectrograms by the scores and generating masks for each source [2]. The score-filtered spectrum is used as input to an encoder-decoder convolutional neural network (CNN) architecture similar to [26]. [27] propose an unsupervised method where scores guide the representation learning to induce structure in the separation. They add class activity penalties and structured dropout extensions to the encoder-decoder architecture. Class activity penalties capture the uncertainty about the target label value and structured dropout uses labels to enforce a specific structure, canceling activity related to unwanted note.

2.2 Text-informed music source separation

Due to the importance of singing voice in a musical piece [12], it is one of the most useful source to separate in a music track. Researchers have integrated the vocal activity information to constrain a robust principal component analysis (RPCA) method, applying a vocal/non-vocal mask or ideal time-frequency binary mask [28]. [10] propose a bidirectional recurrent neural networks (BRNN) method that includes context information extracted from the text via attention mechanism. The method takes as input a whole audio track and its associated text information and learn alignment between mixture and context information that enhance the separation. Recently, [11] extract a representation of the linguistic content related to cognitively relevant features such as phonemes (but they do not explicitly predict the phonemes) in the mixture. The linguistic content guide the synthetization of the vocals.

3. FORMALIZATION

We use the multimodal information as context to guide and improve the separation. We formalize our problem satisfying certain properties summarized as [29]:

How is the multimodal model constructed? We divide the model into two distinct parts [30]: a *generic* network that carries on the main computation and a *control mechanism* that conditions the computation regarding context information and adds additional flexibility. The *conditioning* itself is performed using FiLM layers [17]. FiLM can effectively modulate a generic source separation model by some external information, controlling a single model to perform different instrument source separations [3, 5]. With this strategy, we can explore the *control* and *conditioning* parts regardless of the *generic* network used.

Where is the context information used? at which place in the *generic* network we insert the context information, and defining how it affects the computation, i.e. weak (or strong) conditioning without (or with) explicitly informing where it occurs in the signal.

What context information? We explore here prior information about the phonetic evolution of the singing voice, aligned in time with the audio. To this end, we introduce a novel multitrack dataset with lyrics aligned in time.

4. DATASET

The DALI (Dataset of Aligned Lyric Information) [31] dataset is a collection of songs described as a sequence of time-aligned lyrics. Time-aligned lyrics are described at four levels of granularity: **notes**, **words**, **lines** and **paragraphs**:

$$A_g = (a_{k,g})_{k=1}^{K_g} \text{ where } a_{k,g} = (t_k^0, t_k^1, f_k, l_k, i_k)_g \quad (1)$$

where g is the granularity level and K_g the number of elements of the aligned sequence, t_k^0 and t_k^1 being a text segment's start and end times (in seconds) with $t_k^0 < t_k^1$, f_k a tuple (f_{min}, f_{max}) with the frequency range (in Hz) covered by all the notes in the segment (at the note level $f_{min} = f_{max}$, a vocal note), l_k the actual lyric's information and $i_k = j$ the index that links an annotation $a_{k,g}$ with its corresponding upper granularity level annotation $a_{j,g+1}$. The text segment's events for a song are ordered and non-overlapping - that is, $t_k^1 \leq t_{k+1}^0 \forall k$.

There is a subset of DALI of 513 multitracks with the *mixture* and its separation in two sources, *vocals* and *accompaniment*. This subset comes from WASABI

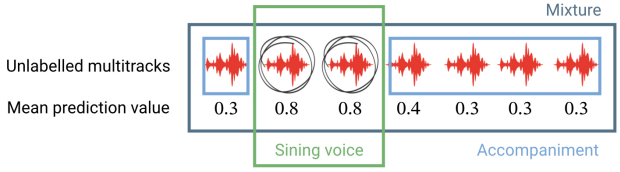

Figure 1. Method used for creating the *vocals*, *accompaniment* and *mixture* version.

dataset [32]. The multitracks are distributed in 247 different artists and 32 different genres. The dataset contains 35.4 hours with music and 14.1 hours with vocals, with a mean average duration per song of 220.83s and 98.97s with vocals. All the songs are in English.

The original multitracks have the *mixture* decomposed in a set of unlabeled sources in the form track_1, track_2, ..., track_n. Depending of the songs, the files can be RAW (where each source is an instrument track e.g. a drum snare) or STEMS (where all the RAW files for an instrument are merged into a single file). In the following, we explain how the *vocals* and *accompaniment* tracks are automatically created from these unlabelled sources. The process is summarized in Figure 1.

For each track τ of a multitrack song, we compute a singing voice probability vector overtime, using a pre-trained Singing Voice Detection (SVD) model [33]. We obtain then a global mean prediction value per tracks ϵ_τ. Assuming that there is at least one track with vocals, we create the *vocals* source by merging all the tracks with $\epsilon_\tau >= \max_\tau(\epsilon_\tau) \cdot \nu$ where ν is a tolerance value set to 0.98. All the remaining tracks are fused to define the *accompaniment*. We manually checked the resulting sources. The dataset is available at `https://zenodo.org/record/3970189`.

The second version of DALI adds the phonetic information computed for the word level [33]. This level has the words of the lyrics transcribed into a vocabulary of 39 different phoneme symbols as defined in the Carnegie Mellon Pronouncing Dictionary (CMUdict) [1]. After selecting the desired time resolution, we can derive a time frame based phoneme context activation matrix Z, which is a binary matrix that indicates the phoneme activation over time. We add an extra row with the 'non-phoneme' activation with 1 at time frames with no phoneme activation and 0 otherwise. Figure 2 illustrates the final activation matrix.

Although we work only with phonemes per word information, we can derive similar activation matrices for other context information such as notes or characters.

5. METHODOLOGY

Our method adapts the C-U-Net architecture [5] to the singing voice separation task, exploring how to use the prior knowledge defined by the phonemes to improve the vocals separation.

Input representations. Let $X \in \mathbb{R}^{T \times M}$ be the magnitude of the Short-Time Fourier Transform (STFT) with

[1] https://github.com/cmusphinx/cmudict

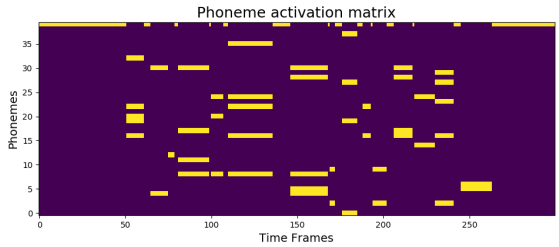

Figure 2. Binary phoneme activation matrix. Note how words are represented as a bag of simultaneous phonemes.

$M = 512$ frequency bands and T time frames. We compute the STFT on an audio signal down-sampled at 8192 Hz using a window size of 1024 samples and a hop size of 768 samples. Let $Z \in \mathbb{R}^{T \times P}$ be the aligned phoneme activation matrix with $P = 40$ phoneme types and T the same time frames as in X. Our model takes as inputs two submatrix $x \in \mathbb{R}^{N \times M}$ and $z \in \mathbb{R}^{N \times P}$ of $N = 128$ frames (11 seconds) derived from X and Z.

Model. The C-U-Net model has two components (see [5] for a general overview of the architecture): a **conditioned network** that processes x and a **control mechanism** that conditions the computation with respect to z. We denote by $x_d \in \mathbb{R}^{W \times H \times C}$ the intermediate features of the **conditioned network**, at a particular depth d in the architecture. W and H represent the 'time' and 'frequency' dimension and C the number of feature channels (or feature maps). A *FiLM* layer conditions the network computation by applying an affine transformation to x_d:

$$FiLM(x_d) = \gamma_d(z) \odot x_d + \beta_d(z) \tag{2}$$

where $\odot$ denotes the element-wise multiplication and $\gamma_d(z)$ and $\beta_d(z)$ are learnable parameters with respect to the input context z. A *FiLM* layer can be inserted at any depth of the original model and its output has the same dimension as the x_d input, i.e. $\in \mathbb{R}^{W \times H \times C}$. To perform Eqn (2), $\gamma_d(z)$ and $\beta_d(z)$ must have the same dimensionaly as x_d, i.e. $\in \mathbb{R}^{W \times H \times C}$. However, we can define them omitting some dimensions. This results in a non-matching dimensionality with x_d, solved by broadcasting (repeating) the existing information to the missing dimensions.

As in [5, 18, 34], we use the U-Net [18] as **conditioned network**, which has an encoder-decoder mirror architecture based on CNN blocks with skip connections between layers at the same hierarchical level in the encoder and decoder. Each convolutional block in the encoder halves the size of the input and doubles the number of channels. The decoder is made of a stack of transposed convolutional operation, its output has the same size as the input of the encoder. Following the original C-U-Net architecture, we insert the *FiLM* layers at each encoding block after the batch normalization and before the Leaky ReLU [5].

We explore now the different **control mechanism** we use for conditioning the U-Net.

5.1 Control mechanism for weak conditioning

Weak conditioning refers to the cases where

Model	U-Net	W_{si}	W_{co}	S_a	S_{a*}	S_c	S_{c*}	S_f	S_{f*}	S_s	S_{s*}
θ	$9.83 \cdot 10^6$	$+14,060$	$+2.35 \cdot 10^6$	$+1.97 \cdot 10^6$	$+327,680$	$+80,640$	$+40,960$	$+40,320$	$+640$	$+480$	$+80$

Table 1. Number of parameters (θ) for the different configurations. We indicate increment to the U-Net architecture.

- $\gamma_d(z)$ and $\beta_d(z) \in \mathbb{R}^1$: they are scalar parameters applied independently of the times W, the frequencies H and the channel C dimensions. They depend only on the depth d of the layer within the network [5].

- $\gamma_d(z)$ and $\beta_d(z) \in \mathbb{R}^C$: this is the original configuration proposed by [17] with different parameters for each channel $c \in 1, ..., C$.

We call them *FiLM simple* (W_{si}) and *FiLM complex* (W_{co}) respectively. Note how they apply the same transformation without explicitly informing where it occurs in the signal (same value over the dimension W and H).

Starting from the context matrix $z \in \mathbb{R}^{N \times P}$, we define the **control mechanism** by first apply the autopool layer proposed by [35] [2] to reduce the input matrix to a timeless vector. We then fed this vector into a dense layer and two dense blocks each composed by a dense layer, 50% dropout and batch normalization. For *FiLM simple*, the number of units of the dense layers are 32, 64 and 128. For *FiLM simple*, they are 64, 256 and 1024. All neurons have ReLU activations. The output of the last block is then used to feed two parallel and independent dense layer with linear activation which outputs all the needed $\gamma_d(z)$ and $\beta_d(z)$. While for the *FiLM simple* configuration we only need 12 γ_d and β_d (one γ_d and β_d for each of the 6 different encoding blocks) for the *FiLM complex* we need 2016 (the encoding blocks feature channel dimensions are 16, 32, 64, 128, 256 and 512, which adds up to 1008).

5.2 Control mechanism for strong conditioning

In this section, we extend the original *FiLM* layer mechanism to adapt it to the strong conditioning scenario.

The context information represented in the input matrix z describes the presence of the phonemes $p \in \{1, \ldots, P\}$ over time $n \in \{1, \ldots N\}$. As in the popular Non-Negative Matrix factorization [36] (but without the non-negativity constraint), our idea is to represent this information as the product of tensors: an activation and two basis tensors.

The **activation tensor** z_d indicates which phoneme occurs at which time: $z_d \in \mathbb{R}^{W \times P}$ where W is the dimension which represents the time at the current layer d (we therefore need to map the time range of z to the one of the layer d) and P the number of phonemes.

The **two basis tensors** γ_d **and** $\beta_d \in \mathbb{R}^{H \times C \times P}$ where H is the dimension which represents the frequencies at the current layer d, C the number of input channels and P the number of phonemes. In other words, each phoneme p is represented by a matrix in $\mathbb{R}^{H \times C}$ derived from Eqn (1). This matrix represents the specific conditioning to apply

[2] The auto-pool layer is a tuned soft-max pooling that automatically adapts the pooling behavior to interpolate between mean and max-pooling for each dimension

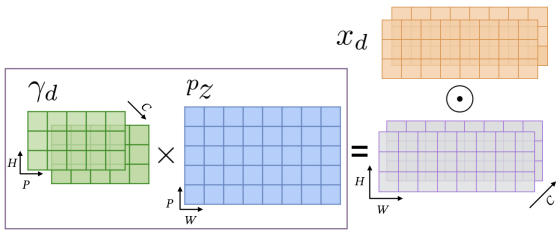

Figure 3. Strong conditioning example with $(\gamma_d \times z_d) \odot x_d$. The phoneme activation z_d defines how the basis tensors (γ_d) are employed for performing the conditioning on x_d.

to x_d if the phoneme exists (see Figure 3). These matrices are learnable parameters (neurons with linear activations) but they do not depend on any particular input information (at a depth d they do not depend on x nor z), they are rather "activated" by z_d at specific times. As for the 'weak'conditionning, we can define different versions of the tensors

- the **all-version** (S_a) described so far with three dimensions: $\gamma_d, \beta_d \in \mathbb{R}^{H \times C \times P}$

- the **channel-version** (S_c): each phoneme is represented by a vector over input channels (therefore constant over frequencies): $\gamma_d, \beta_d \in \mathbb{R}^{C \times P}$

- the **frequency-version** (S_f): each phoneme is represented by a vector over input frequencies (therefore constant over channels): $\gamma_d, \beta_d \in \mathbb{R}^{H \times P}$

- the **scalar-version** (S_s): each phoneme is represented as a scalar (therefore constant over frequencies and channels): $\gamma_d, \beta_d \in \mathbb{R}^P$

The global conditioning mechanism can then be written as

$$FiLM(x_d, z_d) = (\gamma_d \times z_d) \odot x_d + (\beta_d \times z_d) \quad (3)$$

where $\odot$ is the element-wise multiplication and $\times$ the matrix multiplication. We broadcast γ_d and β_d for missing dimensions and transpose them properly to perform the matrix multiplication. We test two different configurations: inserting FiLM at each encoder block as suggested in [5] and inserting FiLM only at the last encoder block as proposed at [3]. We call the former 'complete' and the latter 'bottleneck' (denoted with $*$ after the model acronym). We resume the different configurations at Table 1.

6. EXPERIMENTS

DATA. We split DALI into three sets according to the normalized agreement score η presented in [31] (see Table 2). This score provides a global indication of the global alignment correlation between the annotations and the vocal activity.

	Train	Val	Test
Threshold	$.88 > \eta >= .7$	$.89 > \eta >= .88$	$.89 > \eta$
Songs	357	30	101

Table 2. DALI split according to agreement score η.

Training	Test	Aug	SDR	SIR	SAR
Musdb18	Musdb18	False	4.27	13.17	5.17
(90)	(50)	True	4.46	12.62	5.29
	Musdb18	False	4.60	14.03	5.39
DALI	(50)	True	4.96	13.50	5.92
(357)	DALI	False	3.98	12.05	4.91
	(101)	True	4.05	11.40	5.32

Table 3. Data augmentation experiment.

DETAILS. We train the model using batches of 128 spectrograms randomly drawn from the training set with 1024 batches per epoch. The loss function is the mean absolute error between the predicted vocals (masked input mixture) and the original vocals. We use a learning rate of 0.001 and the reduction on plateau and early stopping callbacks evaluated on the validation set, using patience of 15 or 30 respectively and a min delta variation for early stopping to $1e-5$. Our output is a Time/Frequency mask to be applied to the magnitude of the input STFT mixture. We use the phase of the input STFT mixture to reconstruct the waveform with the inverse STFT algorithm.

For the strong conditioning, we apply a `softmax` on the input phoneme matrix z over the phoneme dimension P to constrain the outputs to sum to 1, meaning it lies on a simplex, which helps in the optimization.

6.1 Evaluation metrics

We evaluate the performances of the separation using the mir evaltoolbox [37]. We compute three metrics: Source-to-Interference Ratios (SIR), Source-to-Artifact Ratios (SAR), and Source-to-Distortion Ratios (SDR) [38]. In practice, SIR measures the interference from other sources, SAR the algorithmic artifacts introduce in the process and SDR resumes the overall performance. We obtain them globally for the whole track. However, these metrics are ill-defined for silent sources and targets. Hence, we compute also the Predicted Energy at Silence (PES) and Energy at Predicted Silence (EPS) scores [10]. PES is the mean of the energy in the predictions at those frames with silent target and EPS is the opposite, the mean of the target energy of all frames with silent prediction and non-silent target. For numerical stability, in our implementation, we add a small constant $\epsilon = 10^{-9}$ which results in a lower boundary of the metrics to be -80 dB [3]. We consider as silent segments those that have a total sum of less than -25 dB of the maximum absolute in the audio. We report the **median** values of these metrics over the all tracks in the DALI test set. For SIR, SAR, and SDR larger values indicate better performance, for PES and EPS smaller values, mean better performance.

Model	SDR	SIR	SAR	PES	EPS
U-Net	4.05	11.40	5.32	-42.44	-64.84
W_{si}	**4.24**	**11.78**	5.38	**-49.44**	-65.47
W_{co}	**4.24**	**12.72**	5.15	**-59.53**	-63.46
S_a	4.04	**12.14**	5.13	**-59.68**	-61.73
S_{a*}	**4.27**	**12.42**	5.26	**-54.16**	-64.56
S_c	**4.36**	**12.47**	5.34	**-57.11**	-65.48
S_{c*}	**4.32**	**12.86**	5.15	**-54.27**	**-66.35**
S_f	4.10	11.40	5.24	47.75	-62.76
S_{f*}	4.21	**13.13**	5.05	**-48.75**	**-72.40**
S_s	**4.45**	**11.87**	**5.52**	**-51.76**	-63.44
S_{s*}	**4.26**	**12.80**	5.25	**-57.37**	-65.62

Table 4. Median performance in dB of the different models on the DALI test set. In bold are the results that significantly improve over the U-Net ($p < 0.001$) and inside the circles the best results for each metric.

6.2 Data augmentation

Similarly as proposed in [39], we randomly created 'fake' input mixtures every 4 real mixtures. In non-augmented training, we employ the mixture as input and the vocals as a target. However, this does not make use of the accompaniment (which is only employed during evaluation). We can integrate it creating 'fake' inputs by automatically mixing (mixing meaning simply adding) the target vocals to a random sample accompaniment from our training set.

We test the data augmentation process using the standard U-Net architecture to see whether it improves the performance (see Table 3). We train two models on DALI and Musdb18 dataset [40] [3] . This data augmentation enables models to achieve better SDR and SAR but lower SIR. Our best results (4.96 db SDR) are not state-of-the-art where the best-performing models on Musdb18 achieve (approximately 6.60 db SDR) [41].

This technique does not reflect a large improvement when the model trained on DALI is tested on DALI . However, when this model is tested on Musdb18, it shows a better generalization (we have not seen any song of Musidb18 during training) than the model without data augmentation (we gain 0.36 dB). One possible explanation for not having a large improvement on DALI testset is the larger size of the test set. It also can be due to the fact that vocal targets in DALI still contain leaks such as low volume music accompaniment that come from the singer headphones. We adopt this technique for training all the following models.

Finally, we confirmed a common belief that training with a large dataset and clean separated sources improves the separation over a small dataset [42]. Both models trained on DALI (with and without augmentation) improve the results obtained with the models trained on Musdb18.

Since we cannot test the conditioning versions on Musdb (no aligned lyrics), the results on the DALI test (4.05 dB SDR) serves as a baseline to measure the contribution of the conditioning techniques (our main interest).

[3] We use 10 songs of the training set for the early stopping and reduction on plateau callbacks

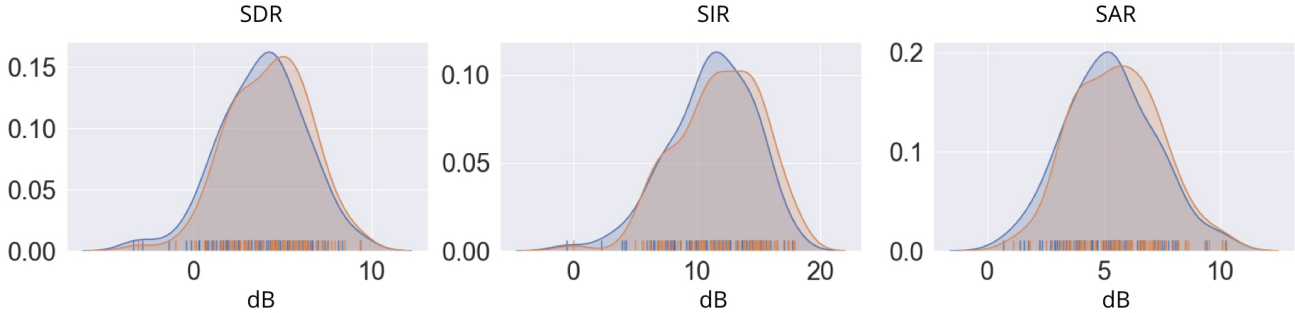

Figure 4. Distribution of scores for the the standar U-Net (Blue) and S_s (Orange).

7. RESULTS

We report the median source separation metrics (SDR, SAR, SIR, PES, ESP) in Table 4. To measure the significance of the improvement differences, we performed a paired t-test between each conditioning model and the standard U-Net architecture, the baseline. This test measures (p-value) if the differences could have happened by chance. A low p-value indicates that data did not occur by chance. As expected, there is a marginal (but statistical significance) improvement over most of the proposed methods, with a generalized $p < 0.001$ for the SDR, SIR, and PES, except for the versions where the basis tensors have a 'frequency' H dimension. This is an expected result since when singing, the same phoneme can be sung at different frequencies (appearing at many frequency positions in the feature maps). Hence, these versions have difficulties to find generic basis tensors. This also explains why the 'bottleneck' versions (for both S_{f*} and S_{a*}) outperforms the 'complete' while this is not the case for the other versions. Most versions also improve the performance on silent vocal frames with a much lower PES. However, there is no difference in predicting silence at the right time (same EPS). The only metric that does not consistently improve is SAR, which measures the algorithmic artifacts introduced in the process. Our conditioning mechanisms can not reduce the artifacts that seem more dependent on the quality of the training examples (it is the metric with higher improvement in the data augmentation experiment Table 3). Figure 4 shows a comparison with the distribution of SDR, SIR, and SAR for the best model S_s and the U-Net. We can see how the distributions move toward higher values.

One relevant remark is the fact that we can effectively control the network with just a few parameters. S_s just adds 480 (or just 80 for S_{s*}) new learnable parameters and have significantly better performance than S_a that adds $1.97 \cdot 10^6$. We believe that the more complex control mechanisms tend to find complex basis tensors that do not generalize well. In our case, it is more effective to perform a simple global transformation. In the case of weak conditioning, both models behave similarly although W_{si} has $1.955 \cdot 10^6$ fewer parameters than W_{co}. This seems to indicate that controlling channels is not particularly relevant.

Regarding the different types of conditioning, when repeating the paired t-test between weak and strong models only S_s outperforms the weak systems. We believe that strong conditioning can lead to higher improvements but several issues need to be addressed. First, there are misalignments in the annotations that force the system to perform unnecessary operations which damages the computation. This is one of the possible explanations of why models with fewer parameters perform better. They are forced to find more generic conditions. The weak conditioning models are robust to these problems since they process z and compute an optimal modification for a whole input patch (11s). We also need to "disambiguate" the phonemes inside words since they occur as a bag of phonemes at the same time (no individual onsets per phonemes inside one word, see Figure 2). This prevents strong conditioning models to properly learn the phonemes in isolation, instead, they consider them jointly with the other phonemes.

8. CONCLUSIONS

The goal of this paper is twofold. First, to introduce a new multimodal multitrack dataset with lyrics aligned in time. Second, to improve singing voice separation using the prior knowledge defined by the phonetic characteristics. We use the phoneme activation as side information and show that it helps in the separation.

In future works, we intend to use other prior aligned knowledge such as vocal notes or characters also defined in DALI . Regarding the conditioning approach and since it is transparent to the conditioned network, we are determined to explore recent state-of-the-art source separation methods such as Conv-Tasnet [43]. The current formalization of the two basis tensors γ_d and β_d does not depend on any external factor. A way to exploit a more complex control mechanisms is to make these basis tensors dependent on the input mixture x which may add additional flexibility. Finally, we plan to jointly learn how to infer the alignment and perform the separation [44, 45].

The general idea of lyrics-informed source separation leaves room for many possible extensions. The present formalization relies on time-aligned lyrics which is not the real-world scenario. Features similar to the phoneme activation [46, 47] can replace them or be used to align the lyrics as a pre-processing step. This two options adapts the current system to the real-world scenario. These features can also help in properly placing and disambiguating the phonemes of a word to improve the current annotations.

Acknowledgement. This research has received funding from the French National Research Agency under the contract ANR-16-CE23-0017-01 (WASABI project). Implementation available at https://github.com/gabolsgabs/vunet

9. REFERENCES

[1] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] M. Miron, J. Janer Mestres, and E. Gómez Gutiérrez, "Monaural score-informed source separation for classical music using convolutional neural networks," in *Hu X, Cunningham SJ, Turnbull D, Duan Z. ISMIR 2017. 18th International Society for Music Information Retrieval Conference; 2017 Oct 23-27; Suzhou, China.[Canada]: ISMIR; 2017. p. 55-62.* International Society for Music Information Retrieval (ISMIR), 2017.

[3] O. Slizovskaia, G. Haro, and E. Gómez, "Conditioned source separation for music instrument performances," *arXiv preprint arXiv:2004.03873*, 2020.

[4] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[5] G. Meseguer-Brocal and G. Peeters, "Conditioned-u-net: Introducing a control mechanism in the u-net for multiple source separations," in *Proc. of ISMIR (International Society for Music Information Retrieval)*, Delft, Netherlands, 2019.

[6] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. Ellis, "Improving universal sound separation using sound classification," *arXiv preprint arXiv:1911.07951*, 2019.

[7] O. Slizovskaia, L. Kim, G. Haro, and E. Gomez, "End-to-end sound source separation conditioned on instrument labels," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 306–310.

[8] P. Seetharaman, G. Wichern, S. Venkataramani, and J. Le Roux, "Class-conditional embeddings for music source separation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 301–305.

[9] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," *arXiv preprint arXiv:2002.07016*, 2020.

[10] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau, "Weakly informed audio source separation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 273–277.

[11] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "Content based singing voice extraction from a musical mixture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 781–785.

[12] A. Demetriou, A. Jansson, A. Kumar, and R. M. Bittner, "Vocals in music matter: the relevance of vocals in the minds of listeners." in *In Proceedings of 19th International Society for Music Information Retrieval Conference*, Paris, France, September 2018.

[13] E. Gómez, M. Blaauw, J. Bonada, P. Chandna, and H. Cuesta, "Deep learning for singing processing: Achievements, challenges and impact on singers and listeners," *arXiv preprint arXiv:1807.03046*, 2018.

[14] E. J. Humphrey, S. Reddy, P. Seetharaman, A. Kumar, R. M. Bittner, A. Demetriou, S. Gulati, A. Jansson, T. Jehan, B. Lehner *et al.*, "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 82–94, 2018.

[15] J. Goldsmith, "Autosegmental phonology," Ph.D. dissertation, MIT Press London, 1976.

[16] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[17] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. of AAAI (Conference on Artificial Intelligence)*, New Orleans, LA, USA, 2018.

[18] A. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proc. of ISMIR (International Society for Music Information Retrieval)*, Suzhou, China, 2017.

[19] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 2013, pp. 1–4.

[20] S. Ewert and M. Müller, "Using score-informed constraints for nmf-based source separation," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 129–132.

[21] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[22] F. J. Rodriguez-Serrano, Z. Duan, P. Vera-Candeas, B. Pardo, and J. J. Carabias-Orti, "Online score-informed source separation with adaptive instrument models," *Journal of New Music Research*, vol. 44, no. 2, pp. 83–96, 2015.

[23] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 888–891.

[24] M. Miron, J. J. Carabias Orti, and J. Janer Mestres, "Improving score-informed source separation for classical music through note refinement," in *Müller M, Wiering F, editors. Proceedings of the 16th International Society for Music Information Retrieval (ISMIR) Conference; 2015 Oct 26-30; Málaga, Spain. Canada: International Society for Music Information Retrieval; 2015.* International Society for Music Information Retrieval (ISMIR), 2015.

[25] M. Miron, J. J. Carabias-Orti, J. J. Bosch, E. Gómez, and J. Janer, "Score-informed source separation for multichannel orchestral recordings," *Journal of Electrical and Computer Engineering*, vol. 2016, 2016.

[26] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *Proc. of LVA/ICA (International Conference on Latent Variable Analysis and Signal Separation)*, Grenoble, France, 2017.

[27] S. Ewert and M. B. Sandler, "Structured dropout for weak label and multi-instance learning and its application to score-informed source separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2277–2281.

[28] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and J.-S. R. Jang, "Vocal activity informed singing voice separation with the ikala dataset," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 718–722, 2015.

[29] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[30] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. d. Vries, A. Courville, and Y. Bengio, "Feature-wise transformations," *Distill*, 2018, https://distill.pub/2018/feature-wise-transformations.

[31] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "Dali: a large dataset of synchronised audio, lyrics and notes, automatically created using teacher-student machine learning paradigm." in *In Proceedings of 19th International Society for Music Information Retrieval Conference*, Paris, France, September 2018.

[32] G. Meseguer-Brocal, G. Peeters, G. Pellerin, M. Buffa, E. Cabrio, C. F. Zucker, A. Giboin, I. Mirbel, R. Hennequin, M. Moussallam *et al.*, "Wasabi: A two million song database project with audio and cultural metadata plus webaudio enhanced client applications," in *Web Audio Conference 2017–Collaborative Audio# WAC2017*, 2017.

[33] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "Creating dali, a large dataset of synchronized audio, lyrics, and notes," *Transactions of the International Society for Music Information Retrieval*, 2020.

[34] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. of ISMIR (International Society for Music Information Retrieval)*, Paris, France, 2018.

[35] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[36] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[37] C. Raffel, B. Mcfee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: a transparent implementation of common mir metrics," in *Proc. of ISMIR (International Society for Music Information Retrieval)*, Porto, Portugal, 2014.

[38] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM TASLP (Transactions on Audio Speech and Language Processing)*, vol. 14, no. 4, 2006.

[39] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 261–265.

[40] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," 2017, https://zenodo.org/record/1117372.

[41] 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.

[42] L. Prétet, R. Hennequin, J. Royo-Letelier, and A. Vaglio, "Singing voice separation: A study on training data," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 506–510.

[43] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[44] K. Schulze-Forster, C. S. Doire, G. Richard, and R. Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7274–7278.

[45] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 41–45.

[46] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. d'Alché Buc, "Audio-based detection of explicit content in music," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 526–530.

[47] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 181–185.