SEPARATION OF UNVOICED FRICATIVES IN SINGING VOICE MIXTURES WITH SEMI-SUPERVISED NMF

*Jordi Janer, Ricard Marxer**

Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
jordi.janer@upf.edu, ricard.marxer@upf.edu

ABSTRACT

Separating the singing voice from a musical mixture is a problem widely addressed due to its various applications. However, most approaches do not tackle the separation of unvoiced consonant sounds, which causes a loss of quality in any vocal source separation algorithm, and is especially noticeable for unvoiced fricatives (e.g. /θ/ in **th**ing) due to their energy level and duration. Fricatives are consonants produced by forcing air through a narrow channel made by placing two articulators close together. We propose a method to model and separate unvoiced fricative consonants based on a semisupervised Non-negative Matrix Factorization, in which a set of spectral basis components are learnt from a training excerpt. We implemented this method as an extension of an existing well-known factorization approach for singing voice (SIMM). An objective evaluation shows a small improvement in the separation results. Informal listening tests show a significant increase of intelligibility in the isolated vocals.

1. INTRODUCTION

In the context of musical audio source separation, we do not find many references in the literature that address the problem of unvoiced phonemes in singing voice. Usually, removing the unvoiced (e.g. fricative) components of the singing voice in a polyphonic mixture is addressed as a joint problem in the signal modeling step. For example, NMF approaches that use harmonic basis sometimes integrate a flat spectrum component to capture the unvoiced parts of the lead vocals [1]. While not specifically addressing singing voice separation, the technique by Wong et al. [2] performs spectral subtraction to obtain the enhanced vocal signal. Then a multilayer perceptron (MLP) is used to segregate the vocal from the non-vocal segments taking as input the spectral flux, the HC, the ZCR, the MFCCs, the amplitude level and the 4Hz modulation energy. Last, the DTW algorithm is used to align the two sequences.

In Hsu and Jang [3], the authors specifically address the problem of unvoiced singing voice separation. A first step segments the signal into accompaniment/voiced/unvoiced predominant frames by means of an HMM. In this case 39 MFCC features were used, computed directly from the STFT (taking energy and the first and second derivatives of the cepstral coefficients). A second step uses a GMM classifier to perform an "Unvoiced-Dominant T-F Unit Identification" within only the unvoiced frames. T-F units are computed by means of a gammatone filter-bank of 128 channels. In the training stage, each T-F unit is labelled as unvoiced-dominant or accompaniment-dominant. This approach seems promising after listening to the results. One drawback seems the large number of parameters to learn (39 features x 32 GMM components x 128 channels), which requires a large amount of training data. However, in addition to audio examples they do provide a publicly available dataset [1].

Recent work has shown that a semi-supervised variation of the NMF can be useful for detecting and modeling individual phonemes in speech. Schmidt and Olsson [4] approach the speech separation problem by using semisupervised sparse NMF. The basis components are previously learned from training data in this technique. The authors show an improvement in the separation when the basis components learned are phoneme-specific. Raj et al.[5] proposed a phoneme-dependent, exemplar-based NMF modeling for speech enhancement of monoaural mixes. The authors created a set of basis components for each phoneme by drawing spectral vectors from segments of speech recordings that contained the target phoneme.

Lately NMF constraints have been widely used in music separation tasks. Various authors[6, 7, 8] have proposed harmonicity and monophonicity constraints by initializing to zero basis or gain bins where the target source is known to have a low or no energy contribution. Ewert and Müller[9] used musical scores to apply harmonicity, fundamental frequency and note-start constraints on the basis and gains of an NMF decomposition of the mixture spectrum. Note attacks were modelled using wideband learned basis components with the gains initialized to 0 at all time frames except those where the notes started.

We propose a method to detect and suppress unvoiced fricative consonants of the singing voice in music mixture recordings. The method extends SIMM with semi-supervised NMF and additional constraints on the factors in order to take into account unvoiced fricative consonants during the singing voice separation. This technique serves as a proof of concept and could be extended to other singing voice components as well as unvoiced components of other musical instruments. The method is tested on a dataset of multitrack music recordings and shows an improvement on the objective perceptual-based separation measures.

2. SPECTRUM MODEL AND FACTORIZATION

The Smoothed Instantaneous Mixture Model (SIMM) introduced by Durrieu et al.[1] is of special interest to us for its flexibility and simplicity. Most of the work presented here can be regarded as an extension of it.

* Authors thank Yamaha Corp. for their support. This research was partially funded by the PHENICX project (EU FP7 Nr. 601166).

[1] http://sites.google.com/site/unvoicedsoundseparation

SIMM is an iterative parameter estimation approach, based on NMF exploiting a source/filter model for the predominant instrument. The code implementing it is available online[2].

This method approximates the mixture magnitude spectrum $\boldsymbol{V}$ as the sum of the lead singing voice and the accompaniment spectra $\hat{\boldsymbol{V}} = \hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_m$. These components are further factorized. The accompaniment is modeled as the non-negative combination of a set of $N_{W\,m}$ constant basis components $\hat{\boldsymbol{X}}_m = \boldsymbol{W}_m \boldsymbol{H}_m$. The singing voice spectrum is approximated as an element-wise multiplication ($\otimes$) of a smooth filter and a monophonic harmonic excitation $\hat{\boldsymbol{X}}_v = \boldsymbol{X}_\Phi \otimes \boldsymbol{X}_{f_0}$. The factor corresponding to the filter is modeled as a combination of constant spectral shapes that are smooth in frequency $\boldsymbol{X}_\Phi = \boldsymbol{W}_\Phi \boldsymbol{H}_\Phi$. To ensure smoothness, the spectral shapes $\boldsymbol{W}_\Phi$ is modeled as a non-negative linear combination of band-limited spectra $\boldsymbol{W}_\Phi = \boldsymbol{W}_\Gamma \boldsymbol{H}_\Gamma$. The monophonicity of the excitation is achieved by modeling it as a non-negative combination of harmonic spectral templates $\boldsymbol{X}_{f_0} = \boldsymbol{W}_{f_0} \boldsymbol{H}_{f_0}$, where all the gains $\boldsymbol{H}_{f_0}$, except a region limited in frequency around the estimated predominant pitch f_0, are set to 0. In this study, we use a low-latency method with timbre classification for singing voice [10] to estimate the predominant pitch f_0.

Some of the presented components are constant. $\boldsymbol{W}_{f_0}$ is composed of harmonic spectra with a magnitude decay computed using the Klatt glottal model. $\boldsymbol{W}_\Gamma$ is a set of band-limited filters, modeled with gaussians centered at frequencies distributed uniformly on the spectrum. The author derives a set of multiplicative update rules to estimate iteratively the other lead-voice components $\boldsymbol{H}_{f_0}$, $\boldsymbol{H}_\Phi$, $\boldsymbol{H}_\Gamma$, as well as the accompaniment components $\boldsymbol{H}_m$ and $\boldsymbol{W}_m$.

3. PROPOSED METHOD

We propose an extension to SIMM that approximates the fricative consonants as an additive wideband component to the singing voice spectra. Using the same notation as the SIMM spectrum model results in $\hat{\boldsymbol{V}} = (\hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_{fric}) + \hat{\boldsymbol{X}}_m$, where $\hat{\boldsymbol{X}}_{fric}$ is interpreted as the spectrum of the fricative consonants of the singing voice, and $\hat{\boldsymbol{X}}_{v'} = \hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_{fric}$ is the full spectrum of the singing voice comprising voiced and fricative components. Similarly to what is done for the accompaniment spectra $\hat{\boldsymbol{X}}_m$ in SIMM, we use an NMF decomposition to model the fricative spectra $\hat{\boldsymbol{X}}_{fric} = \boldsymbol{W}_{fric} \boldsymbol{H}_{fric}$. However in this case the $\boldsymbol{W}_{fric}$ are learned during a training stage and set constant during the separation stage.

The proposed method contains two steps: 1) training a model of NMF basis; and 2) separating the fricatives with the learned NMF basis. To train the NMF basis we provide a recording that contains only the target sounds to separate. In this case we record a sequence of several unvoiced (voiceless) fricative sounds (/s/, /f/, /ʃ/, /θ/, /h/, /tʃ/) by a single subject using a Shure SM-58 microphone (see Figure 1). We apply a low-shelf filter with cutoff frequency at 200Hz to remove the blowing noise (low frequency). We refer to the resulting processed waveform as $x^t_{fric}[t]$.

We then specify the number of basis components $N_{W\,fric}$ to be learned and perform an NMF decomposition of the training audio spectra $\boldsymbol{X}^t_{fric} = \boldsymbol{W}^t_{fric} \boldsymbol{H}^t_{fric}$ into a set of $N_{W\,fric}$ basis components and the corresponding gains $\boldsymbol{H}^t_{fric}$. Both the basis components $\boldsymbol{W}^t_{fric}$ and the gains $\boldsymbol{H}^t_{fric}$ are learned from the data.

[2] `http://durrieu.ch/phd/software.html` (last accessed on January 3, 2011)

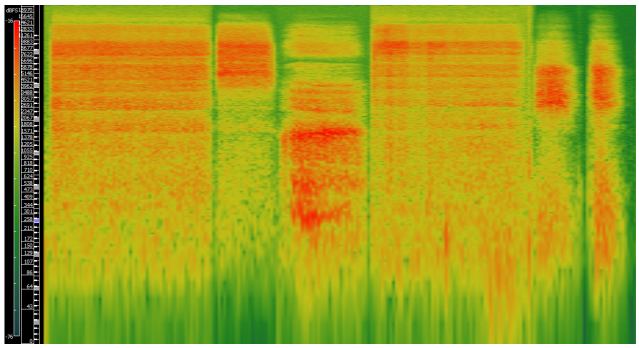

Figure 1: Spectrogram of the six unvoiced fricative sounds used in the NMF training stage (/s/, /f/, /ʃ/, /θ/, /h/, /tʃ/). Frequency y-axis shown in a logarithmic scale.

The resulting $\boldsymbol{H}^t_{fric}$ can be disregarded, since they are only applicable to the specific training input spectrum $\boldsymbol{X}^t_{fric}$. However we can assume $\boldsymbol{W}^t_{fric}$ to be a good generic basis for general vocal fricative instances.

Since the fricative spectra are assumed additive and independent from other factors, the multiplicative update rules are trivial to derive. The update rules of all the components except $\boldsymbol{H}_{fric}$ are computed in the same manner as for SIMM. We must take into account that in the proposed method the estimation of the mixture spectrum $\hat{\boldsymbol{V}}$ also includes the estimated fricative spectra $\hat{\boldsymbol{X}}_{fric}$. The multiplicative update rules for the $\boldsymbol{H}_{fric}$ become:

$$\boldsymbol{H}_{fric} \leftarrow \boldsymbol{H}_{fric} \otimes \frac{{\boldsymbol{W}^t_{fric}}^\top \left(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V} \right)}{{\boldsymbol{W}^t_{fric}}^\top \hat{\boldsymbol{V}}^{(\beta-1)}} \tag{1}$$

Applying the multiplicative rules for a given number of iterations, we obtain the estimated gains of the fricatives $\hat{\boldsymbol{H}}_{fric}$.

The separation of the voice is then done by performing a Wiener filter where the target source is composed of the voiced and fricatives spectra:

$$\boldsymbol{m}_{v'} = \frac{\hat{\boldsymbol{X}}_{v'}}{\hat{\boldsymbol{X}}_{v'} + \hat{\boldsymbol{X}}_m} \tag{2}$$

where $\hat{\boldsymbol{X}}_{v'}[\omega] = \boldsymbol{X}_v[\omega] + \boldsymbol{X}_{fric}[\omega]$ is the estimated vocal source spectrum. The mask is then applied to the complex spectrum of the mixture to compute the estimated source complex spectrum $\widetilde{\hat{\boldsymbol{X}}}_{v'} = \boldsymbol{m}_{v'} \otimes \widetilde{\hat{\boldsymbol{V}}}$. Then a simple overlap-add technique is used to achieve the output waveform signal.

After initial examination of the results, we realize that the main drawback of this approach for estimating the spectra of the fricatives is the use of the fricatives basis components to reconstruct other wideband sources such as hi hats, cymbals or even snare drums. This is due to the similarity of the spectra of these sources. The main difference between these two sources is the transient nature of the sounds. Drums generally create sounds with a very fast increase in energy, which are referred to as transient sounds. On the other hand, fricatives are usually more sustained, with a very slow onset and termination.

In order to overcome this problem we propose using the transient quality of the spectrum frames to differentiate between the fricatives and the drums. Using the same transient estimation method

presented in Janer et al. [11] we extract from the audio mixture a set of N_J transient timepoints t_j^{tr} for $j \in [1 \dots N_J]$.

We assume that at transient positions the fricative presence will be negligeable compared to other sources such as drums or other attacks. These timepoints are then used as constraints on the gains of the fricatives by initializing the corresponding columns to zero:

$$\boldsymbol{H}_{fric}^{T}[w,t] = \begin{cases} 0, & \text{if } |t - t_j^{tf}| < \tau \forall j \\ \gamma, & \text{else} \end{cases} \quad (3)$$

where $\gamma > 0$ is a random positive value and τ is a parameter that controls the size of the vicinity of the transient timepoint.

Following a similar rationale we assume unvoiced fricatives will not be present at the same instants as the pitched singing voice component. Therefore we can define a different initialization based on the estimated singing voice pitch $p[t]$. By defining an unvoiced frame with $p[t] <= 0$ we can determine the initital-ization constraint based on pitch as:

$$\boldsymbol{H}_{fric}^{P}[w,t] = \begin{cases} 0, & \text{if } p[t] > 0 \\ \gamma, & \text{else} \end{cases} \quad (4)$$

Finally we propose another initialization constraint that combines the two previous ones:

$$\boldsymbol{H}_{fric}^{PT}[w,t] = \begin{cases} 0, & \text{if } p[t] > 0 \text{ or } |t - t_j^{tf}| < \tau \\ \gamma, & \text{else} \end{cases} \quad (5)$$

Each of these initialization constraints will lead to a different factorization and separation result. From now on, we will use the names FRIC-T, FRIC-P and FRIC-PT respectively for these separation methods. The method that does not apply any constraint on $\boldsymbol{H}_{fric}$ will be called FRIC.

4. EXPERIMENT

We hypothesize that in the context of singing voice separation the use of trained basis components with transient and pitch-based gains constraints can improve the estimation of unvoiced fricative consonants with the SIMM model. In the experimental setting we test this hypothesis by evaluating the separation results of the SIMM method with the proposed extensions (FRIC, FRIC-T, FRIC-P and FRIC-PT) and without extensions (NONFRIC).

The evaluation material is a dataset of 14 multitrack recordings with vocals, compiled from publicly available resources (MASS[3], SiSEC[4], BSS Oracle[5]) and 2 in-house multitrack recordings.

A quantitative evaluation is done by using the perceptually motivated objective measures in the PEASS toolbox [12]: OPS (Overall Perceptual Score), TPS (Target-related Perceptual Score), IPS (Interference-related Perceptual Score), APS (Artifact-related Perceptual Score).

For all the excerpts we have also computed the near-optimal time-frequency mask-based separation using the BSS Oracle framework. The evaluation measures of the oracle versions of each excerpt were used as references to reduce dependance of the performance on the difficulty of each audio. Therefore the values shown are error values with respect to the near-optimal version. Hence the results shown in the following section, shall be regarded as the lower the better. The main parameter of the proposed method has been set empirically to $\tau = 75$ms for all the tests.

[3] `http://www.mtg.upf.edu/static/mass`
[4] `http://sisec.wiki.irisa.fr/`
[5] `http://bass-db.gforge.inria.fr/bss_oracle/`

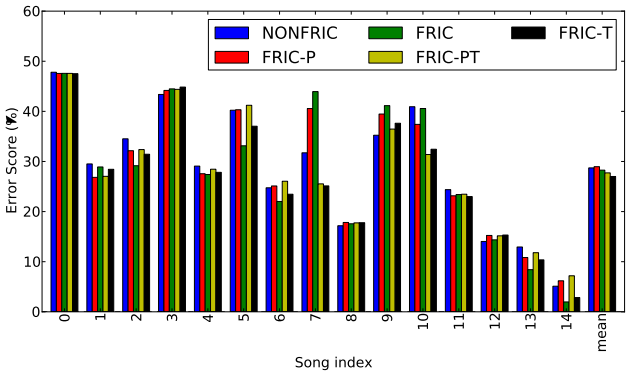

Figure 2: OPS error measures of individual audio examples (indexes in the x-axis). Each columns group is the result of various methods (form left to right): NONFRIC, FRIC-P, FRIC, FRIC-PT and FRIC-T.

5. RESULTS

As shown in Figure 2, results are better on the Overal Perceptual Score in the singing voice isolation task with most of the fricatives estimation methods. This improvement is present in almost all the excerpts, and in those where the proposed methods do not improve, the penalty on the OPS is relatively small. The FRIC-P method, which uses the pitch as a constraint for the fricative gains is the only method that does not improve over the no fricative estimation (NONFRIC).

In Figure 3 and in Table 1 we can observe that the overall separation improvement is mainly due to a decrease in interferences, and the consequent reduction of the IPS error. We also note that the different constraint methods (FRIC-T, FRIC-P, FRIC-PT) have a large influence on the score errors. The use of pitch-based constraints on the fricative gains degrades significantly the separation performance in terms of TPS and APS. Listening examination of the results show that this is mainly due to false positive pitches at fricative positions. Fricatives are often positioned close to voiced phonemes and the pitch estimation used often extends the resulting pitch tracks to these regions due to large analysis windows. Future work could be devoted to studying the adaptation of the pitch estimation and tracking methods to avoid such situations.

Transient-based constraints (FRIC-T and FRIC-PT) improve the overall separation results in comparison to using no constraints at all (FRIC). The improvement of the FRIC-T method comes in the form of a tradeoff between interference and target/artifact errors, possible due to transient constraints being binary thereby leading to non-smooth changes in the time-frequency masks. Informal listening of the results shows mainly a reduction of drums interference in the isolated singing voice, which was the desired effect of the constraint.

We also conducted preliminary tests using basis components trained on "plosive" and "trill" consonants, however the singing voice separation did not improve. Possibly this was due to the lack of characterization of the temporal evolution of their spectra.

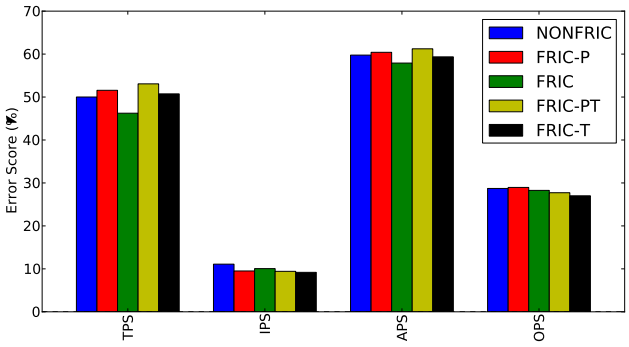

Figure 3: Average error measures for the vocal fricatives separation methods. Each columns group is the result of various methods (form left to right): NONFRIC, FRIC-P, FRIC, FRIC-PT and FRIC-T.

	TPS	**IPS**	**APS**	**OPS**
NONFRIC	50.01	11.10	59.77	28.72
FRIC-P	51.57	9.51	60.42	28.96
FRIC	46.24	10.06	57.91	28.27
FRIC-PT	53.07	9.42	61.23	27.72
FRIC-T	50.74	9.19	59.35	27.01

Table 1: Average error measures for PEASS measures for all the fricative estimation methods.

6. CONCLUSIONS

We have proposed an extension to the SIMM spectrum model and separation technique that takes into account unvoiced fricative consonants when isolating the singing voice. The proposed method makes use of semisupervised NMF to train a set of basis components on audio recordings of isolated fricative consonants, the resulting components are then used in the separation stage. Two types of constraints on the factorization were evaluated. Transient analysis was used to distinguish between percussive events and fricatives. Pitch-based constraints were proposed to restrict the estimation of unvoiced fricatives to regions without pitch. Although the improvement of the objective separation measures is small the perceived difference in informal listening tests is significant. The proposed method is capable of retaining many of the unvoiced fricatives present in the mixture. The transient-based constraints improved the separation by disambiguating between fricatives and drums. However the pitch-based constraints had a negative effect on the separation results, probably due mainly to pitch estimation errors. This research shows the potential of combining semisupervised NMF with model-based factorization such as SIMM. Future work could focus on non-fricative unvoiced consonants such as "plosives" and "trills" to better understand the limitations of the current spectrum model and factorization technique. The use of constraints could be further studied by adapting the pitch estimation techniques to this particular use-case and by testing the methods on ground-truth pitch annotations. The use of regularization could also be an interesting alternative to the constraints, and could reduce the musical artifacts by avoiding the binary masks on the gains matrices.

7. REFERENCES

[1] Jean-Louis Durrieu, Bertrand David, and Gaël Richard, "A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation.," *J. Sel. Topics Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[2] Chi Wong, Wai Szeto, and Kin Wong, "Automatic lyrics alignment for Cantonese popular music," *Multimedia Systems*, vol. 12, no. 4/5, pp. 307–323, Mar. 2007.

[3] Chao-Ling Hsu and J.-S.R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 2, pp. 310–319, feb. 2010.

[4] M.N. Schmidt and R.K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth International Conference on Spoken Language Processing*, 2006.

[5] Bhiksha Raj, Rita Singh, and Tuomas Virtanen, "Phoneme-Dependent NMF for Speech Enhancement in Monaural Mixtures.," in *INTERSPEECH*. 2011, pp. 1217–1220, ISCA.

[6] Toni Heittola, Anssi Klapuri, and Tuomas Virtanen, "Musical Instrument Recognition in Polyphonic Audio Using Source-Filter Model for Sound Separation.," in *ISMIR*, Keiji Hirata, George Tzanetakis, and Kazuyoshi Yoshii, Eds. 2009, pp. 327–332, International Society for Music Information Retrieval.

[7] Jean-Louis Durrieu, Gael Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *IEEE Int. Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, 2009, pp. 105–108.

[8] R. Hennequin, R. Badeau, and B. David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," *Proc. of DAFx-10, Graz, Austria*, pp. 109–112, 2010.

[9] Sebastian Ewert and Meinard Müller, "Score-Informed Voice Separation For Piano Recordings.," in *ISMIR*, Anssi Klapuri and Colby Leider, Eds. 2011, pp. 245–250, University of Miami.

[10] Ricard Marxer, Jordi Janer, and Jordi Bonada, "Low-Latency instrument separation in polyphonic audio using timbre models," *Latent Variable Analysis and Signal Separation*, pp. 314–321, 2012.

[11] J. Janer, R. Marxer, and K. Arimoto, "Combining a harmonic-based NMF decomposition with transient analysis for instantaneous percussion separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 281–284.

[12] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation.," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.