Phoneme-dependent NMF for speech enhancement in monaural mixtures

Bhiksha Raj[1], *Rita Singh*[1], *Tuomas Virtanen*[2]

[1] Carnegie Mellon University, Pittsburgh, PA, USA.
[2] Tampere University of Technology, Tampere, Finland.

Abstract

The problem of separating speech signals out of monaural mixtures (with other non-speech or speech signals) has become increasingly popular in recent times. Among the various solutions proposed, the most popular methods are based on compositional models such as non-negative matrix factorization (NMF) and latent variable models. Although these techniques are highly effective they largely ignore the inherently phonetic nature of speech. In this paper we present a *phoneme-dependent* NMF-based algorithm to separate speech from monaural mixtures. Experiments performed on speech mixed with music indicate that the proposed algorithm can result in significant improvement in separation performance, over conventional NMF-based separation.

Index terms: Monaural signal separation, speech enhancement, restoration, Non-negative matrix factorization.

1. Introduction

The problem of separating signals from monaural mixtures is a difficult one. Given a recording that includes a mixture of sounds from multiple sources, the goal here is to separate out the individual signals, or at least enhance one. Often the problem addressed in literature is that of separating multiple speakers from a monoaural recording, but the techniques are also useful for separating speech out from other background noises.

The particular form of solution to the problem that has become most popular employs *compositional* models, in which the magnitude spectra of signals from any source are modeled as being composed by a constructive linear combination of spectral "bases" that represent characteristic spectro-temporal patterns for that source. The magnitude spectra of *mixed* signals are constructive linear combinations of the bases for all the sources in the mixture. To separate a source from the mixed signal, it is sufficient estimate the contribution of the bases from that source to the mixed signal. The actual compositional model itself can take different forms. The most common one is based on non-negative matrix factorization (NMF) [1]. Schmidt [2] describes the basic NMF-based method that characterizes bases as individual magnitude spectral vectors. Smaragdis [3] presents a convolutive model that employs spectro-temporal *patterns* as bases, rather than simple spectra. Many authors *e.g.* [2, 4] have described *overcomplete* models wherein the number of bases is greater than the dimensionality of spectral representations. Virtanen [5] describes a perceptually-weighted variation that assigns greater importance to separation of low frequencies than to higher ones. Gemmeke *et. al.* [6] embed NMF within an HMM framework to introduce additional temporal constraints on the separation. Other similar solutions exist in the literature. An alternative form of compositional model is the "latent variable" approach described in [7]. This represents magnitude spectra as histograms obtained by draws from a mixture multinomial process. The component multinomials are the bases that characterize the source. Signal separation now becomes a maximum-likelihood or maximum *a-posteriori* estimate of the contribution of the bases from individual sources. Overcomplete representations [4], temporal modeling [8], shift-invariance etc. are now imposed within this structure.

In the compositional framework, a target signal is separated out of a mixed signal by extracting out the contribution of its bases from the mixture. The effectiveness of separation depends critically on having bases that can compose the target signal completely; otherwise the target signal will not be completely extracted. Therefore, to represent all possible variations in the audio from the target source, a large, often overcomplete collection of bases is employed. This overrepresentation of the source causes other problems: the set may include bases that match spectral patterns in the *competing* sources in the mixture. As a result, these unwanted spectral patterns may get erroneously incorporated into the estimate for the signal from the target source.

When the target signal is speech, some fundamental properties of speech could be used to restrict the set of bases employed in the extraction method. Speech signals are composed of *phonemes*. Each phoneme has a distinct spectral structure. Since bases represent spectral structures in the signal, it can be expected that the bases that compose any phoneme will be different from those that compose other phonemes. By ensuring both, that the spectral structures in the segment can be adequately composed by the bases, and that spectral structures not currently in the segment *cannot* be composed, we can avoid the assignment of spectral components from the target speech to the competing source, and minimize the inclusion of spectral patterns from the competing source into the target speech. In this paper we present a *phoneme-constrained* NMF-based signal separation technique that is based on the above principle.

In our method, we learn separate bases for each phoneme of the language. Given a mixed signal of speech and a competing signal, we use the identity of the phoneme in any segment of the speech to select the appropriate speech bases to be used for that segment for separation. Since the phoneme identity is usually unknown, we use an automatic speech recognition (ASR) system to jointly find the optimal phoneme sequence and separate the speech from the competing signal.

In Section 2 below we describe the feature representation used for separation. In Section 3 we briefly recall NMF-based separation of signals. In Section 4 we present our phoneme-dependent NMF-based separation technique. Sections 5 and 6 report our experiments and conclusions.

2. Feature Representation and Notation

We represent all signals as their short-time Fourier transforms, *i.e.* spectrograms. Let $y[t]$ be a mixture of a speech signal $s[t]$

and it's competing signal $n[t]$. The STFT of the mixture can be shown to be the sum of the STFT of component signals:

$$Y(t,f) = S(t,f) + N(t,f) \tag{1}$$

where $Y(t,f)$, $S(t,f)$ and $N(t,f)$ are the values at frequency f in the t^{th} analysis frame of the STFT of $y[t]$, $s[t]$ and $n[t]$ respectively. The additive relation of Equation 1 also approximately holds for the magnitudes of the spectral components, *i.e.* $|Y(t,f)| = |S(t,f)| + |N(t,f)|$. We therefore operate on the magnitude of the STFT of the signals. Accordingly, our separation algorithm too estimates the magnitude spectra of the separated signals. To re-synthesize separated speech, we combine it's estimated magnitude spectrogram with the phase of the spectrogram of the original mixed signal.

In this paper we use the following notation: all references to spectral components refer to their magnitudes, but we omit the standard "|.|" notation for magnitudes. Upper case symbols, *e.g.* Y refer to a *single* magnitude spectral vector. Specific frequency components are indicated as $Y(t)$ or $Y(f)$ ("t" generally refers to time and "f" to frequency). We use the notation $Y(t,f)$ when we use both. Bold upper case characters, *e.g.* $\mathbf{Y}$ represent collections or sequences of magnitude spectral vectors.

3. NMF for Signal Separation

3.1. The Compositional Model

The compositional model represents any magnitude spectral vector S of speech as a weighted linear non-negative combination of speech basis vectors B_i as

$$S = \sum_{i=1}^{N} B_i w_i(S) \tag{2}$$

where B_i is the i^{th} basis vector and $w_i(S)$ is the weight of the basis. N is the number of speech basis vectors. The weight $w_i(S)$ is specific to the vector S – a different vector may need a different weight. The bases B_i are magnitude spectral vectors and are strictly non-negative. The weights $w_i(S)$ too are all non-negative. The intuition behind this is that any sound is composed by constructive composition of its components, *e.g.*, a segment of music may be composed by additive composition of the notes that comprise it. Cancellation, which is represented by negative weights, rarely, if ever, factors into the composition of a sound, except by careful design.

If we represent the complete set of basis vectors as a matrix $\mathbf{B}_s = \left[B_1, \ldots, B_N\right]$, and the weights as a vector $W(S) = \left[w_1(S), \ldots, w_N(S)\right]^\top$, we can write Equation 2 as:

$$S = \mathbf{B}_s W(S) \tag{3}$$

Similarly, the competing signal is modeled as a weighted sum of bases for its source. Representing it's bases a matrix $\mathbf{B}_n$, and the weights with which they must be combined to form a spectral vector N as $W(N)$, the model for any spectral vector N from the competing source can be written as $N = \mathbf{B}_n W(N)$. The model for a mixed spectral vector $Y = S + N$ can now be written as

$$\begin{aligned} Y &= S + N = \mathbf{B}_s W(S) + \mathbf{B}_n W(N) \\ &= \mathbf{B}W \end{aligned} \tag{4}$$

where $\mathbf{B} = [\mathbf{B}_s \mathbf{B}_n]$ is a matrix that combines the bases for speech and competing source into a single matrix, and $W = [W(S)^\top W(N)^\top]^\top$ combines the weights $W(S)$ and $W(N)$ into a single vector.

3.2. The bases

The bases in $\mathbf{B}_s$ and $\mathbf{B}_n$ are also spectral vectors. Equation 2 does not specify how the bases are obtained. This is by choice. In prior work we found *exemplar-based* characterizations [9, 10], that use realizations of spectral vectors from the source signals itself as the bases, to be highly effective for signal separation. We therefore use this method to derive bases. We obtain the speech basis vectors $\mathbf{B}_s$ as magnitude spectral vectors drawn randomly from training examples of speech. Similarly, the bases for the competing source, $\mathbf{B}_n$ are obtained by drawing random spectral vectors from examples of the competing signal.

3.3. Estimating Weights

Once the set of bases $\mathbf{B}$ is given, the vector of weights W with they must be combined to optimally compose Y is estimated using an update rule that minimizes a generalized Kullback-Leibler divergence between Y and the composition $\mathbf{B}W$ [1]. This rule estimates the weights through iterations of:

$$W = W \otimes \frac{\mathbf{B}^\top . [\frac{Y}{\mathbf{B}.W}]}{\mathbf{B}^\top . \mathbf{1}} \tag{5}$$

The operation $\otimes$ represents element-wise multiplication. All divisions too are element- wise. We initialize all the weights W to unity and iterate Equation 5 to convergence. Thereafter, the weight vector for speech, $W(S)$ is obtained by splitting W as $W = [W(S)^\top W(N)^\top]^\top$.

3.4. Signal reconstruction

From the estimated $W(S)$, the minimum-mean-squared-error estimate of S, *i.e.* the contribution of speech to Y can be obtained using the following Wiener filter formulation:

$$\hat{S} = (Y + \epsilon) \otimes \frac{\mathbf{B}W}{\mathbf{B}W + \epsilon} \tag{6}$$

The ϵ is a scalar adjustment factor that determines how much of the residual signal $(Y - \mathbf{B}W)$ is reallocated to $\hat{S}$. Setting this to a large value gives us $\hat{S} = \mathbf{B}_s W(S)$. $\epsilon = 0$ gives us a conventional Wiener filter. The reconstituted speech spectrogram is converted to a time-domain signal by combining it with the phase from the complex spectrogram of the mixed signal, applying an inverse STFT, and overlap-add combination of the frames. This procedure can also be viewed as filtering the mixed signal with a time-varying filter defined by $\mathbf{B}_s W(S)/\mathbf{B}W$, similarly to Wiener filtering.

4. Phoneme dependent separation

Phonemes are basic sound units that compose understandable speech. For instance, in English the sounds $/L/$ and $/AY/$ and $/K/$ that compose the pronunciation of the word "LIKE" are all phonemes. Typically, all instances of a phoneme share similar spectral structure, a feature that allows them to be visually and automatically recognized in spectrograms. An example is shown in Figure 1. Knowing the phoneme for any segment of speech thus enables us to restrict the bases required for it to those that are required to compose instances of that phoneme. This is the basic principle behind phoneme-dependent separation.

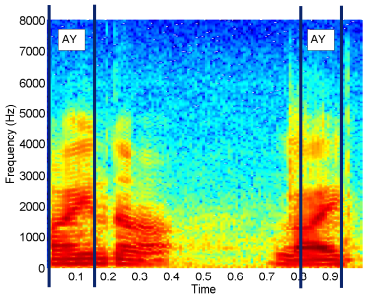

Figure 1: Spectrograms of two instances of the phoneme /AY/, occurring in the words "MINOR" and "LIKE".

4.1. Phoneme-dependent speech bases

To employ phoneme-dependent processing, we need a mechanism for identifying and locating the phonemes in a recording. For this we use an ASR system. For the work reported here, we assume that speakers speak English, and use a 40-phoneme set utilized by CMUDICT (http://www.speech.cs.cmu.edu/cgi-bin/cmudict), an opensource dictionary that lists the phonetic composition of a large number of English words. We learn a separate set of bases $\mathbf{B}_s(/p/)$ for each phoneme $/p/$ in the speech signal. Given a corpus of training speech recordings, we derive a phoneme segmentation for them using an ASR system. To learn the phoneme-specific bases $\mathbf{B}_s(/p/)$, we draw spectral vectors from segments of the training speech that contain $/p/$.

4.2. Separation of Speech from a Mixture

Let $\hat{S} = Sep(Y; \mathbf{B}_s)$ represent the operation by which a mixed signal spectrum $Y = S + N$ is processed using speech bases $\mathbf{B}_s$ to estimate the separated speech signal $\hat{S}$. The function $Sep(Y; \mathbf{B}_s)$ represents the separation procedure described in Section 3. The notation does not explicitly represent the bases for the competing source, for brevity and clarity of expression. Let $/p/(S)$ be the actual phoneme for a spectral vector S. The separation, using phone-specific bases, is now given by:

$$\hat{S} = Sep(Y; \mathbf{B}_s(/p/(S))) \tag{7}$$

If the phoneme identity for the speech in any Y is known *a priori* Equation 7 can directly be employed. More generally, however, it must be estimated. We do so using the following iterative algorithm:

1. In an initial separation step, we use a *generic* set of bases $\mathbf{B}_s$ to separate out $\hat{\mathbf{S}}$ from the $\mathbf{Y}$.
2. We estimate the phoneme sequence for the signal $\hat{s}(t)$ derived from $\hat{\mathbf{S}} = \{\hat{S}\}$ using an ASR system.
3. We use these phoneme labels to perform phoneme-dependent separation to obtain $\hat{\mathbf{S}}$ from the $\mathbf{Y}$.
4. If the procedure has not converged, we return to step 2.

We can state the procedure above as a maximum-likelihood estimation procedure, although we have not done so here. Therein, the convergence criterion employed in Step 4 can be the likelihood assigned to the estimated speech by the ASR system. However, it is also an *unsupervised* method prone to poor local optima. In practice, we have found that after 1 or 2 iterations, the resulting separated spectrograms begin to degrade. Once the separated spectrogram is obtained, it can be converted back to a signal as before.

5. Experimental Evaluation

We conducted experiments on speech that was digitally mixed with music. Speech was taken from the designated training data of the Wall Street Journal (WSJ0) database. For this paper we used recordings from only one speaker (01m). We set aside 10 recordings from the speaker as test data, and used the rest to derive bases. The signals were corrupted using music from the RWC database [11] to a variety of SNRs. RWC includes several hours of recordings of music of various genres. The music corrupting the speech was randomly drawn from the entire database. The STFT employed to parameterize the signals used 40ms windows with a 10ms shift between frames. For all recognition-based results, the Sphinx-III open-source speech recognition system was used. The recognizer was trained using the 1997 broadcast news corpus from LDC, making it independent of the remaining experimental setup.

For the baseline results with generic (phoneme-independent) bases a total of 6000 bases were drawn randomly from all recordings for the speaker. For phoneme-dependent NMF, the training recordings for the speaker were segmented into phonemes automatically, using the recognizer and separate sets of up to bases 1000 bases were obtained for each phoneme (for phonemes such as $/ZH/$, the actual number of vectors available were much fewer, so fewer bases were drawn). Bases for the music were also drawn randomly from the RWC corpus (not including regions used to corrupt signals). The genre of the music corrupting any signal was assumed to be unknown in all experiments.

A "control" test was run to establish that phoneme identities do indeed help improve separation. For this test, the phoneme segmentation for the test utterances were obtained directly from the uncorrupted clean speech. Figure 2 shows an example of the separation obtained. The middle panel of the figure shows the separated spectrogram obtained with generic NMF. The bottom panel shows the separated spectrogram obtained with known phoneme identities. The recording for this example was corrupted to 0dB by music. We note that phoneme-dependent separation using the correct phoneme identity greatly improves the separation of the spectrogram. Similar results were obtained for other SNRs. Additional results may be found at the website listed later in this section.

We note here that the optimal value of ϵ in the reconstruction formula of Equation 6 was different for the baseline (phoneme-independent) and phoneme-dependent separation algorithms. For the baseline, the optimal value was 0. For the phoneme-dependent case, the best *spectrograms* (with the highest correlation with the spectrograms of clean speech) were obtained at high values of ϵ. The resynthesized *speech*, however, sounded best for lower ϵ values (equal to roughly half the maximum amplitude of the spectrogram of the mixed signal). Figure 3 shows examples of the separation obtained when the phoneme segmentation is obtained automatically. The figure shows the results of the first and second iterations of the unsupervised phoneme-dependent separation procedure. The unsupervised separation too gives significantly improved results over phoneme-independent separation (middle panel of 2), although it is not as good as when the phoneme identity is known. Also, the second iteration is actually worse than the first. Generally, the best separation was obtained after one or two iterations; thereafter the separation degraded.

The ASR system used to obtain the phoneme segmentation in Figure 3 performed *all-phone* recognition, where we directly attempt to estimate the phoneme sequence in the record-

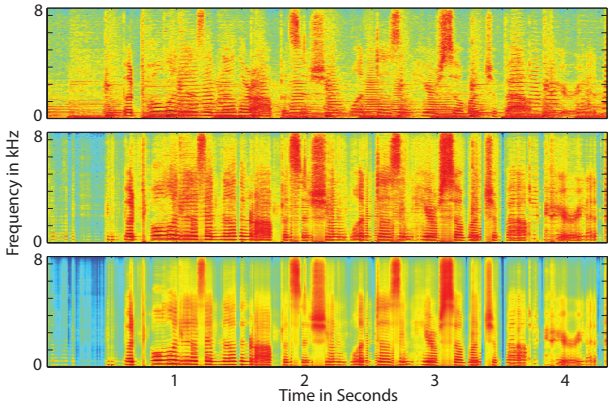

Figure 2: *Top*: Spectrogram of mixed signal. *Middle*: Separation result obtained with conventional NMF. *Bottom*: Result from phone-dependent NMF, when phoneme identities are known *a priori*.

ing. When *word* recognition is performed instead, strong statistical constraints may be derived from a statistical language model. In Figure 4 we show the separtion obtained when we use a word recognizer, and derive phoneme segmentations from the hypothesized word sequence. While not immediately apparent from the figure, this results in inconsistent performance – in regions where the recognizer is correct, or has hypothesized genuine acoustic confusions, the separation is better than that obtained with phoneme recognition. However, word-based recognizers, being forced to recognize words, will sometimes hypothesize acoustically unrelated words. In these portions of the signal, the separation actually degrades, as in the blue patch to the right of Figure 3. These and other results, including audio examples and additional analysis, may be found at our website: http://mlsp.cs.cmu.edu/projects/audio/phonemedependentnmf

6. Conclusion

In our experiments we used music as the "competing signal". The technique, however, is generic and can be extended to deal with speech-over-speech, by hypothesizing phoneme sequences for both signals. Also, all reported experiments were speaker-dependent. However, we demonstrated in [12] that NMF-based separation works just as well when the identity of the speaker is unknown, provided the speaker is represented by the bases; we therefore expect the results to extend to that scenario as well. There are several options going forward, such as the use of perceptual weighting, clustering of phonemes into acoustically confusable classes for more robust separation, and better integration into the recognizer. We will explore these in future work.

7. References

[1] D. D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401(6755), pp. 788–791, 1999.

[2] Mikkel N. Schmidt, *Single-channel source separation using non-negative matrix factorization*, Ph.D. thesis, Technical University of Denmark, Nov 2008.

[3] Paris Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. on audio, speech and language processing*, vol. 15(1), pp. 1–12, 2007.

[4] M. V Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete decomposition for single-channel speaker separation," in *Proc. ICASSP*, 2007.

[5] Tuomas Virtanen, "Monaural sound source separation by perceptually weighted non-negative matrix factorization," Tech. Rep., Tampere University of Technology, 2007.

[6] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. ICASSP*, 2010.

[7] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as non-negative factorizations," *Computational intelligence and Neuroscience*, 2008.

[8] G. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *9th Intl. conf. on latent variable analysis and source separation*, 2010.

[9] P. Smaragdis, R. Shashanka, and B. Raj, "A Sparse Non-Parameteric Approach for Single Channel Separation of Known Sounds," *Proc. NIPS*, 2009.

[10] J. F. Gammecke and T. Virtanen, "Noise-robust exemplar-based connected digit recognition," *Proc. ICASSP*, 2010.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Soun d Database," *International Conference on Music Information Retrieval*, 2003.

[12] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for robust speech recognition," in *Proc. Interspeech*, 2010.

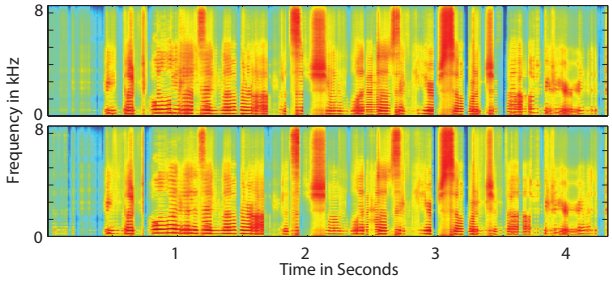

Figure 3: Unsupervised phoneme-dependent separation. *Top*: Spectrogram after one iteration of phone-dependent separation. The initial spectrogram used to obtain phoneme segmentations is the one in the middle panel of Figure 2. *Bottom*: After two iterations.

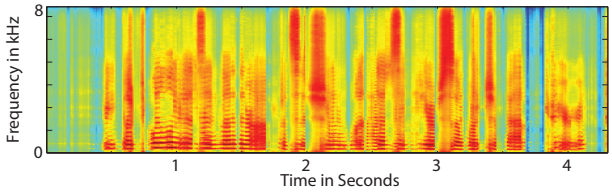

Figure 4: Unsupervised phoneme-dependent separation. Here a word-level recognizer generated the hypotheses and phoneme segmentations were derived from it.