A TWO-STAGE APPROACH FOR IMPROVING THE PERCEPTUAL QUALITY OF SEPARATED SPEECH

Donald S. Williamson[⋆] *Yuxuan Wang*[⋆] *DeLiang Wang*[⋆†]

[⋆] Department of Computer Science and Engineering, The Ohio State University, USA
[†]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{williado,wangyuxu,dwang}@cse.ohio-state.edu

ABSTRACT

Binary time-frequency masking and model-based nonnegative matrix factorization (NMF) are two common approaches to speech separation. However, binary masking often suffers from poor perceptual quality, while NMF typically requires pretrained models for both speech and noise and frequently does not perform well. In this paper we examine whether a single or two-stage approach should be used for performing separation. We propose a two-stage algorithm that uses a soft mask in the first stage for separation, and NMF in the second stage for improving perceptual quality where only a speech model needs to be trained. We show that the proposed two-stage approach achieves higher objective perceptual quality and intelligibility compared to related single-stage methods.

Index Terms— nonnegative matrix factorization, speech separation, speech quality, binary masking

1. INTRODUCTION

Separating speech from noise is a challenging task that has been studied extensively. Robust automatic speech recognition, speaker identification, and hearing prosthesis all benefit from algorithms that successfully perform this task. Commonly used approaches for separating speech from noise include nonnegative matrix factorization (NMF) and binary masking [1, 2, 3, 4].

NMF is a model-based approach that uses trained speech and noise models, along with an activation matrix to separate noisy speech [1, 2, 5, 6]. This approach often requires knowledge of the speaker and noise. A recent improvement to supervised NMF is the nonnegative factorial hidden Markov model (N-FHMM) [7, 8]. This semi-supervised approach uses a nonnegative hidden Markov model (N-HMM) to model speech, while the model for the noise is determined during the separation process. N-HMM uses several small dictionaries and HMM to model the spectral structure and temporal dynamics of speech, respectively. N-FHMM produces a Wiener mask that is used to separate the speech from the noise.

Binary masking often amounts to ideal binary mask (IBM) estimation, i.e. to determine whether a time-frequency (T-F) unit is speech or noise dominant. The resulting binary mask is applied directly to the T-F representation of the noisy speech to obtain a speech estimate. Various techniques exist for estimating the IBM [3, 4]. A common problem with this approach is that the incorrect classification of T-F units leads to a degradation of speech quality.

Both of the above mentioned approaches may be viewed as single stage. We propose to use two stages to perform speech separation, mainly for enhanced speech quality. In the first stage a deep neural network (DNN) is used to produce a soft mask to perform separation. In the second stage we use an NMF basis matrix that is trained just from clean speech to reconstruct the speech separated by the soft mask. Our two-stage approach is compared to several single-stage approaches, including binary masking, supervised NMF, and semi-supervised N-FHMM. Unlike our previous method [9], we use a soft mask instead of a binary mask and an NMF basis matrix instead of a sparse representation approach for the second stage. We evaluate our system using PESQ [10] and STOI [11] to measure perceptual speech quality and predicted intelligibility, respectively.

The rest of the paper is organized as follows. The proposed algorithm is presented in Section 2. Our approach and the comparison approaches are evaluated in Section 3. Section 4 concludes the paper.

2. PROPOSED METHOD

In this section, we describe our proposed two-stage approach. We start with a description of our first stage, which generates a soft mask. We then describe our NMF stage that enhances the quality of the speech separated by the soft mask.

2.1. First stage: soft mask separation

In the first stage of our approach we use DNNs to generate a soft mask that separates speech from background noise. Specifically, the DNNs are trained from the following features that are extracted from the gammatone filter responses

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

of noisy speech training data: amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), and mel-frequency cepstral coefficients (MFCC), as well as their deltas [12]. Separate DNNs are trained for each channel of the 64-channel gammatone filterbank [3], where the IBM is used for ground truth labels. The same features are also extracted from test mixtures, and used along with the trained DNNs to generate a soft mask. The output of each DNN can be interpreted as the posterior probability of a T-F unit being speech dominant. We use this posterior probability as our soft mask. This differs from the approaches in [3, 9], where the output of the DNN is binarized to form a binary mask.

For each test mixture, the soft mask is applied to the gammatone response of the mixture to produce estimated speech. The estimated speech is re-synthesized to the time domain and the STFT of the time-domain signal is computed. Figure 1 shows the spectrogram for a clean speech, noisy speech signal at -5 dB, and the speech estimate that is generated by applying the soft mask to the noisy speech.

After separation, a sliding window augments the STFT magnitude response by combining M frames ($\frac{M-1}{2}$ before and after the current frame, along with the current frame) into a single vector, resulting in a $MN \times T$ matrix [13]. Note that M is an odd integer and N and T correspond to the number of frequency channels and the number of time frames in the STFT, respectively.

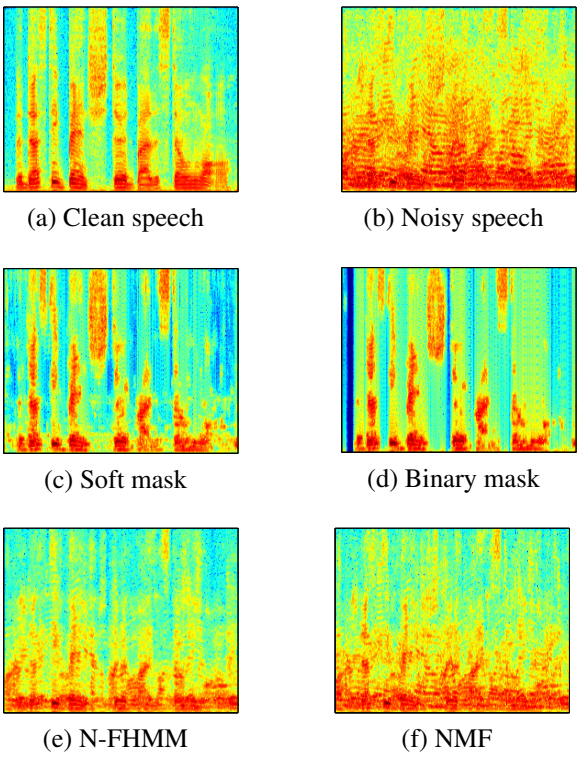
2.2. Second stage: soft mask/NMF reconstruction

NMF uses a linear combination of basis vectors to approximate a signal [5, 6]. We use NMF to enhance the augmented STFT magnitude $\boldsymbol{S}$ of speech separated by a soft mask (i.e. output from Section 2.1). Two nonnegative matrices are used and they are a trained basis matrix $\boldsymbol{W}^{train}$ that defines the spectral features, and an activation matrix $\boldsymbol{H}$ that linearly combines the spectral features.

The basis matrix is trained from clean speech training data $\boldsymbol{D}$, using the approach described in [13], where the product of the trained basis and activation matrices approximates the training data.

$$\boldsymbol{D} \approx \boldsymbol{W}^{train}\boldsymbol{H}^{train} \tag{1}$$

The training data is the concatenation of augmented STFTs from clean speech utterances. The trained basis and activation matrices are determined by minimizing a cost function between the training data and the product of the matrices. There are a multitude of cost functions that are used for NMF, but the generalized Kullback-Leibler (KL) divergence has worked well for source separation [1, 2], thus we use it for the approximation. The empirical mean and covariance of the log values of $\boldsymbol{H}^{train}$, denoted by $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, are also computed and are used during the reconstruction process.



(a) Clean speech (b) Noisy speech (c) Soft mask (d) Binary mask (e) N-FHMM (f) NMF

Fig. 1. Example spectrograms of clean speech, noisy speech and single-stage separation approaches. The noisy speech has a signal-to-noise ratio of -5 dB and is produced using factory noise.

Once $\boldsymbol{W}^{train}$ is computed, $\boldsymbol{S}$ is approximated as the product of the trained basis matrix and a new activation matrix $\boldsymbol{H}$ (i.e. $\boldsymbol{S} \approx \boldsymbol{W}^{train}\boldsymbol{H}$). $\boldsymbol{H}$ is computed using the regularized NMF approach defined in [2]. Specifically, the update rule for each entry in the activation matrix is as follows:

$$H_{ab} \leftarrow H_{ab}\frac{\sum_i W^{train}_{ia} S_{ib}/(\boldsymbol{W}^{train}\boldsymbol{H})_{ib}}{\sum_k W^{train}_{ka} + \beta\,\varphi(\boldsymbol{H})} \tag{2}$$

$$\varphi(\boldsymbol{H}) = -\frac{(\boldsymbol{\Lambda}^{-1}(log(\boldsymbol{H}_{:,b}) - \boldsymbol{\mu}))_a}{H_{ab}} \tag{3}$$

where H_{ab} refers to the element at the a^{th} row and b^{th} column of $\boldsymbol{H}$. The parameter β encourages consistency between the statistics of $\boldsymbol{H}$ and $\boldsymbol{H}^{train}$. Note that in (2) the trained basis matrix is held constant for each iteration.

$\boldsymbol{S}$ is a $MN \times T$ matrix that is converted back to a $N \times T$ matrix by appropriately unwrapping, placing, and averaging the multiple responses within each frame. The estimated STFT magnitude response is combined with the noisy phase response of the mixture to produce an SM/NMF reconstructed STFT. An estimate of the speech signal is produced by performing overlap-and-add synthesis.

3. EXPERIMENTS AND DISCUSSION

We perform speech separation using 60 male utterances randomly selected from the IEEE speech corpus [14], which are downsampled to 12 kHz. Each utterance is mixed with random cuts of 20-talker babble, factory, and speech-shaped noise at -5 and 0 dB, resulting in a total of 360 test examples.

The DNNs are trained by mixing 390 male clean speech utterances randomly selected from the IEEE speech corpus, with random cuts of the noises mentioned above at -5 and 0 dB. The STFT magnitudes from 10 male IEEE clean speech utterances are augmented, concatenated, and used to train the NMF basis matrix. The STFTs are augmented with $M = 5$ frames and are generated using a window and hop size of 20 and 10 ms, respectively. The NMF training basis matrix has 80 basis vectors. The utterances used for testing, training the DNNs, and training the basis matrix are different and do not overlap.

As a comparison of two-stage approaches, we also train a N-HMM from the concatenated STFT magnitudes described previously. A N-HMM uses several small dictionaries (rather than a single large dictionary that is used for NMF) to model the non-stationarity of speech, and a HMM to model temporal dynamics. In a given time frame the speech is modeled as a linear combination of the spectral components from one of the many dictionaries of the N-HMM [7]. As in [7, 8], we use 40 dictionaries of 10 spectral components each to reconstruct the speech separated by the soft mask, using the approach described in [15]. We also use the concatenated STFT magnitudes directly and use the sparse reconstruction approach from [9] as another comparison, which is denoted as EBM/Sparse.

We compare our two-stage approach to binary masking, semi-supervised N-FHMM [8], and supervised NMF [2]. The binary mask is generated by binarizing the output of the DNNs (i.e. the T-F unit is speech dominant if the posterior probability is greater than 0.5), which is done in [9]. The binary mask is applied to the STFT of noisy speech to produce a speech estimate. The N-FHMM models noisy speech using a trained N-HMM speech model, while iteratively learning the parameters for the noise model during speech separation. Each time frame of noisy speech is modeled by a linear combination of the spectral components from the concatentation of one of the speech dictionaries and the noise dictionary. The trained N-HMM mentioned above is used for the speech model of the N-FHMM. The noise is modeled using 1 dictionary of 10 spectral components, as in [8]. The N-FHMM returns a Wiener mask that is applied to the STFT of noisy speech to obtain a speech estimate. The supervised NMF approach uses the above mentioned NMF training basis matrix as a speech model. The noise is modeled by computing a basis matrix from the concatenation of augmented STFT magnitudes from the 20-talker babble, factory, and speech-shaped noises used above. Eighty basis vectors are used for the noise basis matrix. We also performed semi-supervised N-FHMM and supervised NMF using non-augmented STFTs [8, 2], but the performance is better when using augmented STFTs.

	PESQ Score		STOI Score	
	-5 dB	0 dB	-5 dB	0 dB
Noisy Speech	1.3586	1.6189	0.5454	0.6561
Binary Mask	1.4416	1.8194	0.6664	0.7749
Soft Mask	**1.8444**	**2.1628**	**0.7036**	**0.8073**
N-FHMM	1.6535	1.9683	0.5822	0.6967
NMF	1.5128	1.7997	0.5636	0.6780

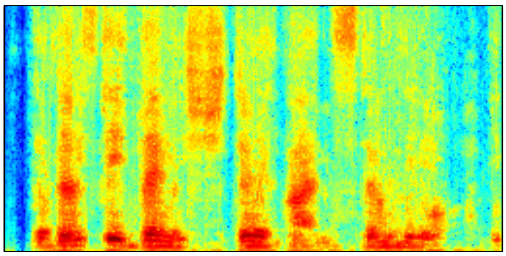
Table 1. Average PESQ and STOI scores for noisy speech and the single-stage approaches.

	PESQ Score		STOI Score	
	-5 dB	0 dB	-5 dB	0 dB
EBM/Sparse	1.6871	2.0389	0.6989	0.7858
SM/NMF	**2.0887**	**2.3596**	**0.7483**	**0.8210**
SM/N-HMM	1.9400	2.1219	0.7457	0.8015

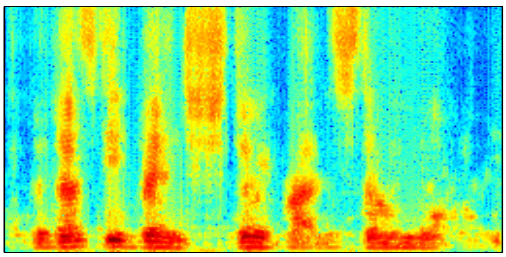
Table 2. Average PESQ and STOI scores for the two-stage approaches. SM/NMF refers to our proposed approach that is described in Section 2.

We use PESQ to evaluate the speech quality of the different approaches. PESQ is an objective perceptual speech quality measure that returns scores between -0.5 and 4.5, where higher scores correspond to higher perceptual speech quality [10]. The predicted intelligibility of the different approaches is evaluated with STOI, which is an objective intelligibility measure [11]. STOI scores range between 0 and 1, where higher scores indicate higher intelligiblity. Table 1 shows the PESQ and STOI scores for noisy speech at -5 and 0 dB, along with results for the single-stage approaches. Notice that our soft mask produces the highest scores over all single-stage approaches in terms of PESQ and STOI at each signal to noise ratio, which justifies its use as the first stage of our two-stage approach. The soft mask produces larger speech quality gains over the binary mask because it does not produce musical noise that is prevalent with binary masks. The soft mask outperforms semi-supervised N-FHMM because a single small dictionary (i.e. a dictionary with 10 spectral components) cannot always explain the subtleties of speech [8], where a speech dictionary with 10 spectral components was found to be optimal in [15]. N-FHMM produces noticeable improvements in terms of PESQ over the noisy speech and the binary mask, however, according to STOI the intelligibility is not much improved over the noisy speech. As expected, N-FHMM outperforms supervised NMF since it models the spectral structure and temporal dynamics of speech. Figure 1 also shows spectrogram examples for these approaches.

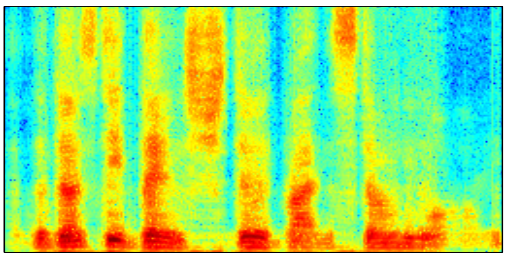
The PESQ and STOI scores of the different two-stage approaches are shown in Table 2. Notice that each approach



(a) Speech separated using EBM/Sparse approach



(b) Speech separated using SM/NMF approach



(c) Speech separated using SM/N-HMM approach

Fig. 2. Example spectrograms for the different two-stage approaches.

provides improvements in terms of PESQ and STOI over its single-stage counterpart (i.e. EBM/Sparse improves binary masking and SM/NMF provides improvements over soft mask) except for SM/N-HMM at 0 dB. Since a speech model can only adequately model speech (and not noise), the second stage is able to suppress some of the noise that remains after the first stage of separation. SM/NMF performs better than SM/N-HMM for the same reason that the soft mask outperforms N-FHMM. On the other hand, SM/NMF performs better than EBM/Sparse because the former approach gets better performance in the first stage with the soft mask. Example spectrograms for the different two-stage approaches are shown in Figure 2. Experiments were also conducted where the individual speech models (i.e. sparse, NMF, and N-HMM) where used to model noisy speech, but no improvements were made over the un-processed noisy speech, indicating that the first stage for separation is necessary.

4. CONCLUSIONS

We have proposed a two-stage approach for improving the perceptual quality of separated speech. In the first stage of our approach, a DNN generates a soft mask that separates speech from background noise. We then reconstruct the speech separated by the soft mask, using nonnegative matrix factorization. This proposed two-stage approach significantly improves perceptual quality and intelligibility, and outperforms single-stage approaches and other two-stage approaches.

5. ACKNOWLEDGMENT

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048), and the Ohio Supercomputer Center. We thank Gautham Mysore for providing assistance to our N-HMM and N-FHMM implementations.

6. REFERENCES

[1] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and spareness criteria," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 15, 2007.

[2] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 4029–4032.

[3] Y. Wang and D. L. Wang, "Towards scaling up classication-based speech separation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 21, pp. 1381–1390, 2013.

[4] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, 2009.

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

[6] H. S. Seung and D. Lee, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[7] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Proc. International Conference on Latent Variable Analysis and Signal Separation*, 2010.

[8] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011, pp. 17–20.

[9] D. S. Williamson, Y. Wang, and D. L. Wang, "A sparse representation approach for perceptual quality improvement of separated speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7015–7019.

[10] ITU-R, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," p. 862, 2001.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, pp. 2125–2136, 2011.

[12] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 21, pp. 270–279, 2013.

[13] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, pp. 2067–2080, 2011.

[14] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[15] G. J. Mysore, *A Non-negative Framework for Joint Modeling of Spectral Structure and Temporal Dynamics in Sound Mixtures*, Ph.D. thesis, Stanford University, USA, 2010.