A SPARSE REPRESENTATION APPROACH FOR PERCEPTUAL QUALITY IMPROVEMENT OF SEPARATED SPEECH

Donald S. Williamson[1] *Yuxuan Wang*[1] *DeLiang Wang*[1,2]

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

ABSTRACT

Speech separation based on time-frequency masking has been shown to improve intelligibility of speech signals corrupted by noise. A perceived weakness of binary masking is the quality of separated speech. In this paper, an approach for improving the perceptual quality of separated speech from binary masking is proposed. Our approach consists of two stages, where a binary mask is generated in the first stage that effectively performs speech separation. In the second stage, a sparse-representation approach is used to represent the separated signal by a linear combination of Short-time Fourier Transform (STFT) magnitudes that are generated from a clean speech dictionary. Overlap-and-add synthesis is then used to generate an estimate of the speech signal. The performance of the proposed approach is evaluated with the Perceptual Evaluation of Speech Quality (PESQ), which is a standard objective speech quality measure. The proposed algorithm offers considerable improvements in speech quality over binary-masked noisy speech and other reconstruction approaches.

Index Terms— Sparse Representations, Speech Quality, Binary Masking, Ideal Binary Mask (IBM), Speech Separation

1. INTRODUCTION

Computational Auditory Scene Analysis (CASA) systems have been used extensively to separate speech signals that are corrupted by noise in a monaural recording. A main computational goal of CASA is to estimate the ideal binary mask (IBM) that identifies whether a time-frequency (T-F) unit is dominated by speech or noise [1]. A T-F unit is assigned a value of 1 if it is speech dominant, and 0 otherwise. An estimate of the speech signal is then obtained by applying the binary mask to the T-F representation of the mixture.

When applying a binary mask to the T-F representation of a mixture, portions of the target speech are removed when they are considered to be dominated by noise. Likewise, portions of the noise are retained when they are considered to be dominated by speech. This creates a problem in speech quality, which is typically evaluated by comparing the estimated speech signal against the clean speech signal [2, 3, 4]. Also, errors in binary mask estimation degrade perceptual speech quality due to musical noise and cross-talk problems [5].

Methods have been proposed to address the quality issue in speech separation. In particular, [5] attempts to reduce the effects of musical noise by smoothing the binary mask in the cepstral domain. In [2], musical noise is reduced by using a fine shift rate when generating T-F representations. These approaches reduce the effects of musical noise, however, the effects due to incorrectly defining a speech dominant T-F unit as noise dominant is not addressed. In [6], a hybrid approach that combines a model-based approach with a source-driven approach is used to separate speech in 0 dB monaural recordings. A multi-pitch tracker extracts pitch information from the speakers, while the model-based approach uses a vector quantizer to represent the spectrum envelope of the speakers. Although the hybrid approach results in improved signal-to-noise ratios (SNRs) over the individual approaches, using pitch may not be completely effective for improving speech quality since estimating pitch at low SNRs is very challenging [4].

In this paper, we propose to use a sparse representation technique to reconstruct the STFT magnitudes of speech separated by binary masking. With sparse representations, each time frame of the STFT magnitude of separated speech is replaced by a sparse linear combination of STFT magnitudes from clean speech. Sparse representations have been effective in similar tasks such as automatic speech recognition (ASR) [7] and image denoising [8, 9], however, its ability to improve the perceptual quality of speech separated by binary masking has not been investigated. The proposed approach utilizes sparse representations to improve the perceptual quality of separated speech, and our system will be compared against traditional STFT magnitude reconstruction approaches. PESQ will be used as the objective speech quality measure to evaluate our systems performance.

The rest of the paper is organized as follows. The proposed algorithm is presented in Section 2. An evaluation of our approach is given in Section 3, along with a comparison to other reconstruction approaches. Section 4 concludes the

discussion of the proposed system.

2. DESCRIPTION

Our proposed system is a two-stage approach, where initially a binary mask is estimated that identifies the T-F units that are speech dominant and the T-F units that are noise dominant. The binary mask is applied to the STFT of the mixture to produce estimated STFTs for the speech and the noise, respectively. A ratio mask is generated from the STFT magnitudes of the separated speech and noise. The ratio mask is then applied to the STFT of the original noisy mixture, resulting in a new estimated STFT. Using sparse representations, the STFT magnitude of the speech signal separated by the ratio mask is represented as a sparse linear combination of STFT magnitudes from clean speech signals. Finally, the sparsely-reconstructed STFT magnitude is combined with the STFT phase of the mixture, and overlap-and-add synthesis is used to produce an estimated speech signal. The following sections describe these steps in more detail.

2.1. IBM Estimation

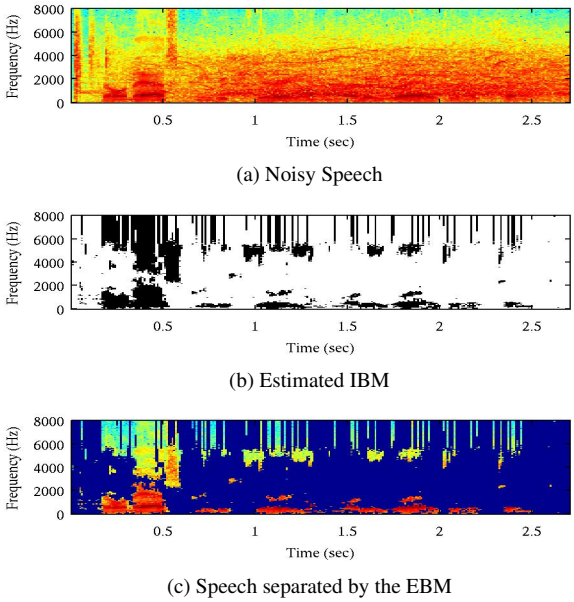
The IBM is computed from the STFT magnitudes of the speech component, $S(t,f)$, and the noise component, $N(t,f)$, of a mixture, where t and f index the time and frequency dimensions, respectively. With these two representations, the IBM is a binary matrix defined as follows [1]:

$$IBM(t,f) = \begin{cases} 1, & \text{if } S(t,f) > N(t,f) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The complement of the binary mask assigns a value of 1 if it is noise dominant and 0 otherwise. An estimate of the noise is obtained when the complementary binary mask is applied to the T-F representation of the mixture.

In our proposed approach, a binary mask is generated by binary classification (e.g., [10]). A set of complementary features such as amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), mel-frequency cepstral coefficients (MFCC), pitch-based, and delta features, are extracted from the input mixture. Using these features, a deep neural network (DNN) generates a binary mask by classifying whether a T-F unit is speech or noise dominant. Unlike [10], temporal dynamics is not used to generate a binary mask. Also, the binary mask returned by the DNN is in the gammatone domain, and it is subsequently converted to the STFT domain for our proposed approach.

Figure 1 shows the spectrogram for a noisy speech signal at an SNR of 0 dB, the estimated binary mask (EBM), and the spectrogram resulting from applying the EBM to the noisy speech STFT. Notice that the EBM spectrogram is incomplete where many of the T-F units have been completely removed.



(a) Noisy Speech

(b) Estimated IBM

(c) Speech separated by the EBM

Fig. 1. Spectrograms for the noisy speech signal (a) and the speech signal separated by the EBM (c). The EBM is shown in (b), where black indicates that the T-F unit is speech dominant, and white indicates it's noise dominant.

Estimated speech and noise STFT magnitudes ($\hat{S}(t,f)$ and $\hat{N}(t,f)$) are generated by applying the gammatone-domain binary mask and its complement to the STFT of the mixture, followed by overlap and add synthesis to produce speech and noise estimates. A ratio mask, $RM(t,f)$, is generated by using the STFT magnitudes of the speech and noise estimates:

$$RM(t,f) = \frac{\hat{S}(t,f)}{\hat{S}(t,f) + \hat{N}(t,f)} \quad (2)$$

A new STFT magnitude for the separated speech is generated by applying the ratio mask to the STFT magnitude of the mixture. A ratio mask is used over a binary mask so that the resulting STFT is complete.

2.2. Sparse Representations of STFT Magnitude

The STFT magnitude of the separated speech inevitably contains noise elements that may negatively affect the perceptual quality of the separated speech signal. To combat this effect, a sparse representation approach is used to denoise the STFT magnitudes.

The underlying principle behind sparse representations is that a given signal can be represented as a sparse linear combination of basis vectors [8]. The signal in our domain is the STFT magnitude $\mathbf{X}$. Given a dictionary (or basis) $\mathbf{D} \in \mathbb{R}^{N \times K}$, the STFT magnitude $\mathbf{X} \in \mathbb{R}^{N \times T}$ is approximated as $\mathbf{D}*\mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{K \times T}$. Letting $\mathbf{A}=[\hat{\boldsymbol{\alpha}}_1,\hat{\boldsymbol{\alpha}}_2,\cdots,\hat{\boldsymbol{\alpha}}_T]$, $\hat{\boldsymbol{\alpha}}_t \in \mathbb{R}^K$ defines the sparse basis vectors from the dictionary, $\boldsymbol{d}_k$'s, that

are used to represent $\mathbf{x}_t$, the t^{th} time frame of $\mathbf{X}$ for $1 \leq k \leq K$ and $1 \leq t \leq T$, (3). Typically, the dictionary $\mathbf{D}$ is underdetermined so $K \gg N$.

$$\begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_T \end{bmatrix} \approx \begin{bmatrix} \boldsymbol{d}_1 & \cdots & \boldsymbol{d}_K \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_1 & \cdots & \hat{\boldsymbol{\alpha}}_T \end{bmatrix} \quad (3)$$

The important steps in sparse representations are then to define the dictionary $\mathbf{D}$, compute the parameters $\boldsymbol{\alpha}_t$ for a given STFT magnitude $\mathbf{X}$, and determine the number of basis vectors, L, that combine to provide an estimate of $\mathbf{X}$ (the STFT magnitude of the separated speech). The dictionary $\mathbf{D}$ is generated by the concatenation of STFT magnitudes of clean speech utterances. Given $\mathbf{D}$ and $\mathbf{X}$, $\mathbf{A}$ is found by solving the following equation,

$$\hat{\boldsymbol{\alpha}}_t = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 \;\; s.t. \;\; \|\boldsymbol{\alpha}\|_1 \leq L, \; 1 \leq t \leq T \quad (4)$$

where L is a parameter that controls sparseness. Since the goal is to approximate the STFT magnitude of speech separated by the ratio mask using sparse representations, the only constraint for L is that it is much smaller than the number of vectors in the dictionary, i.e., $L \ll K$.

With the parameters solved, the STFT magnitude of the separated speech signal $\mathbf{X}$ is approximated by $\hat{\mathbf{X}} \approx \mathbf{D}*\mathbf{A}$. Unlike $\mathbf{X}$, $\hat{\mathbf{X}}$ no longer contains noisy T-F units. The approximated magnitude response $\hat{\mathbf{X}}$ is combined with the noisy-phase information from the mixture, to produce a sparsely-reconstructed STFT. An estimate of the speech signal is then produced by performing the overlap-and-add synthesis on the sparsely-reconstructed STFT.

3. EVALUATIONS AND COMPARISONS

The proposed system is evaluated by employing 100 clean speech utterances randomly selected from the TIMIT corpus. Each utterance is approximately 2 to 4-seconds long, and is sampled at 16 kHz. Each of the speech utterances are mixed with 10 non-speech noises at a SNR of 0 dB, resulting in a test set of 1000 mixtures.

The STFT is computed for each mixture. This is accomplished by windowing the mixture with a sequence of overlapping 20 ms Hamming windows, and then computing the Fast-Fourier Transform (FFT) of the windowed signal. An overlap amount of 50% is used between adjacent frames.

In order to perform sparse representations for the STFT magnitudes, a dictionary must first be generated. The dictionary is generated by concatenating the STFT magnitudes of clean speech utterances from 110 speakers selected from the TIMIT corpus. This results in 1000 utterances used to train the dictionary for sparse representation. Preliminary tests indicated that the number of basis vectors to sparsely represent a signal should be set to 5 (i.e., L = 5). Also note that the testing and training sets are disjoint.

	PESQ score	STOI score
Mixture	1.93	0.76
EBM	1.63	0.72
IBM	3.07	0.92

Table 1. Average PESQ and STOI scores for the noisy speech signals, speech separated by the EBM, and speech separated by the IBM.

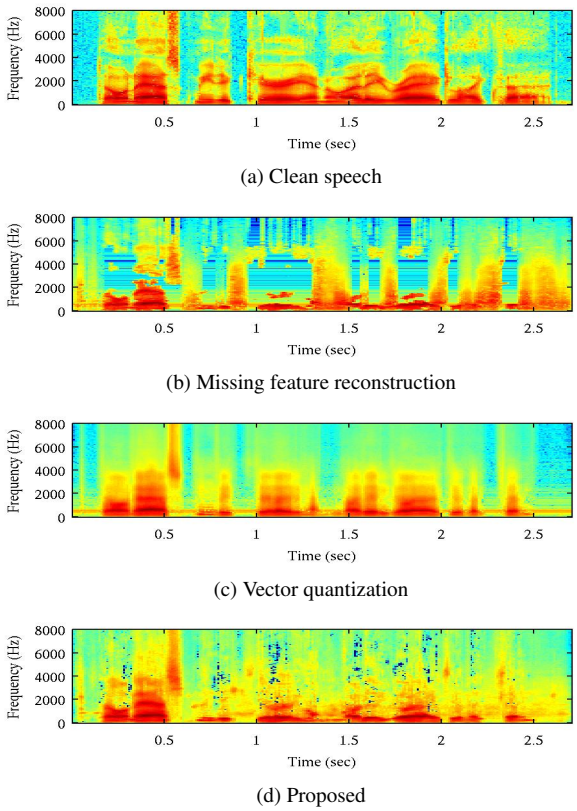
	PESQ Score		STOI Score	
	EBM	IBM	EBM	IBM
Reconstruction [13]	1.69	2.76	0.63	0.87
VQ	0.90	2.12	0.42	0.80
Proposed	**2.29**	**2.84**	**0.78**	**0.89**

Table 2. Average PESQ and STOI scores for different STFT magnitude reconstruction approaches, when EBMs and IBMs are used, respectively.

The speech quality of the sparsely-reconstructed speech signals are evaluated by PESQ, which is an objective perceptual speech quality measure [11]. PESQ scores are between -0.5 and 4.5, where higher scores correspond to higher perceptual speech quality. A PESQ score is computed by comparing the clean speech signal of the mixture against a degraded signal (i.e., the output signal after sparse representations). This is possible because we have access to the pre-mixed clean signals for each test mixture.

To show the effectiveness of sparse representations for denoising STFT magnitudes, we compare our system against two other STFT magnitude reconstruction approaches, namely missing feature reconstruction [12, 13] and vector quantization (VQ) that is based on the approach in [6]. For the missing feature reconstruction approach [12, 13], a speaker-adapted Gaussian Mixture Model (GMM) Universal Background Model (UBM) is trained from the same 1000 utterances that were used to train the dictionary for sparse representations [14]. The GMM was modeled with 64 Gaussians and diagonal covariance matrices. The codebook for the VQ approach was also trained with the same 1000 utterances, but in this case 1024 codewords were used. With missing feature reconstruction, T-F units that are classified as noise dominant by the EBM are replaced with estimated values that are based on the speech-dominant T-F units. In the VQ approach, each time frame from the STFT magnitude of speech separated by the ratio mask is replaced by the closest codeword from the codebook, where closeness is measured in terms of mean square error.

Table 1 shows the average PESQ score for the baseline signals, namely the unprocessed mixture, the signal resulting from applying the EBM without any further processing, and the signal resulting from applying the IBM without any further processing. Notice that applying the EBM without additional STFT magnitude reconstruction results in a lowering



(a) Clean speech

(b) Missing feature reconstruction

(c) Vector quantization

(d) Proposed

Fig. 2. Spectrograms for the clean speech signal (a), missing feature reconstruction approach (b), vector quantization approach (c), and the proposed sparse representation approach (d). The EBM is used for binary masking in each reconstruction approach

of PESQ score as compared to the original unprocessed mixture, while the IBM results in a substantial improvement of speech quality. These results indicate the need for a second stage to improve the PESQ performance from the EBM. The average short-time objective intelligibility (STOI) scores for the baseline signals are also shown in Table 1. STOI is a standard objective measure that quantifies the intelligibility of a signal [15]. STOI scores are between 0 and 1, where higher scores indicate higher intelligibility.

The STOI and PESQ scores of the different reconstruction approaches are shown in Table 2. Notice that the PESQ scores using missing feature reconstruction and VQ are significantly worse than the PESQ scores of the unprocessed mixtures when the EBM is used as the binary mask. However, our proposed sparse representation approach leads to a significant PESQ improvement over the unprocessed mixtures. The proposed approach considerably outperforms the standalone EBM and the other reconstruction approaches. Our proposed approach also results in a significant improvement in objective intelligibility compared to the other reconstruction approaches when the EBM is used as the binary mask, as indicated by the STOI scores. The proposed approach also improves STOI performance over the unprocessed mixture slightly, unlike the other reconstruction approaches. It is worth noting that our approach generates significant PESQ improvements over binary masking using EBM, and does so without degrading STOI scores - it actually yields an improvement from 0.72 to 0.78.

The sparse representation approach also provides the best PESQ and STOI performances, compared to the other STFT magnitude reconstruction approaches, when the IBM is used for binary masking. However, all of the reconstruction approaches lower the PESQ and STOI performances when compared to the IBM with no additional processing. This may occur because the IBM correctly identifies the T-F units that are speech dominant and noise dominant.

Example spectrograms for each of the reconstruction approaches are shown in Figure 2, where the EBM was used for binary masking in each reconstruction approach. Compared to the spectrogram of the clean speech signal, the spectrogram of the proposed approach appears to be most similar. The spectrogram of the missing feature reconstruction approach contains a substantial amount of differences from the clean speech; for example a considerable amount of noise is added around the 2.5 second mark. Likewise, the result from the VQ approach tends to smooth out the harmonics of the clean speech utterance. Although the spectrogram from our approach appears closest to the clean speech spectrogram, it still appears to remove some of the harmonic information from the voiced frames, and the harmonics that are present are not as distinct as in the clean speech case. Thus, there is clearly room for improvements.

4. CONCLUSION

In this paper, a novel approach for improving the perceptual quality of speech separated by a binary mask has been proposed. In this approach, an estimated binary mask is initially determined by using a DNN classifier. We estimate a new STFT magnitude by using the property that signals can be represented as a sparse linear combination of basis vectors. This proposed approach significantly improves the perceptual quality of separated speech, and outperforms other reconstruction approaches. The intelligibility of the separated speech is also improved according to an objective intelligibility measure. To our knowledge, this is the first study that uses a sparse representation to denoise STFT magnitudes and improve perceptual quality of speech separated with binary masks.

5. ACKNOWLEDGMENT

This research was supported by an AFOSR grant (FA9550-12-1-0130) and a NIDCD grant (R01 DC012048). We also thank Xiaojia Zhao and Yang Shao for their implementation.

6. REFERENCES

[1] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. 2005, pp. 181–197, Kluwer.

[2] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proc. ICASSP*, 2005, vol. 3, pp. iii/81–iii/84.

[3] D.L. Wang, "Time–frequency masking for speech separation and its potential for hearing aid design," *Trends in amplification*, vol. 12, pp. 332–353, 2008.

[4] P. Mowlaee, R. Saeidi, M. G. Christensen, Zheng-Hua Tan, T. Kinnunen, P. Franti, and S. H. Jensen, "A joint approach for single-channel speaker identification and speech separation," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 20, pp. 2586 –2601, 2012.

[5] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proc. ICASSP*, 2008, pp. 45–48.

[6] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Monaural speech segregation based on fusion of source-driven with model-driven techniques," *Speech Communication*, vol. 49, pp. 464 – 476, 2007.

[7] J. Gemmeke and B. Cranen, "Using sparse representations for missing data imputation in noise robust speech recognition," *in Proc. EUSIPCO*, 2008.

[8] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Proc.*, vol. 15, no. 12, pp. 3736 –3745, 2006.

[9] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Proc.*, vol. 17, pp. 53 –69, 2008.

[10] Y. Wang and D. L. Wang, "Towards scaling up classication-based speech separation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 2013.

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

[12] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.

[13] X. Zhao, Y. Shao, and D.L. Wang, "CASA-based robust speaker identification," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 20, pp. 1608 –1616, 2012.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, pp. 19–41, 2000.

[15] 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.