DEEP NEURAL NETWORKS FOR ESTIMATING SPEECH MODEL ACTIVATIONS

*Donald S. Williamson, Yuxuan Wang, and 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

This paper presents an approach for improving the perceptual quality of speech separated from background noise at low signal-to-noise ratios. Our approach uses two stages of deep neural networks, where the first stage estimates the ideal ratio mask that separates speech from noise, and the second stage maps the ratio-masked speech to the clean speech activation matrices that are used for nonnegative matrix factorization (NMF). Supervised NMF systems make assumptions about the relationship between the activation and basic matrices that do not always hold. Other two-stage approaches combining masking with NMF reconstruction do not account for mask estimation errors. We show that the proposed algorithm achieves higher objective speech quality and intelligibility compared to these related methods.

Index Terms— nonnegative matrix factorization, deep neural network, speech quality, speech separation

1. INTRODUCTION

Nonnegative matrix factorization (NMF) has been used in many algorithms for separating speech from background noise. NMF approximates a signal as the product between a basis matrix and an activation matrix, where the basis matrix provides spectral structure and the activation matrix linearly combines the basis matrix elements [1, 2]. The main goal of NMF is to train an appropriate basis matrix that provides a generalized spectral representation of speech and to determine an activation matrix that successfully combines the basis elements, so that the error between the signal and its approximation is minimized.

Supervised NMF uses trained speech and noise models that when linearly combined estimate noisy speech [3, 4, 5, 6]. Its objective during speech enhancement is to produce an activation matrix that is split into a speech portion and a noise portion, where the portions are combined with the corresponding model to generate an approximation of the speech and noise components of the mixture. For instance, the first portion of the activation matrix when combined with the speech basis matrix approximates the speech portion of the mixture, whereas the later portion of the activation matrix and the noise model estimates the noise. An assumption is made that the speech model and activations only approximate the speech portion of the signal and do not provide any approximations for the noise. This assumption does not always hold, however, since portions of the speech model and activations may approximate the noise source, especially at low signal-to-noise ratios (SNRs) or when the noise exhibits speech-like qualities.

In [7, 8], it is shown that combining a masking approach with NMF produces higher quality speech than supervised NMF approaches and other two-stage methods. A deep neural network is used to estimate a time-frequency (T-F) mask that when applied to the noisy mixture produces a speech estimate. This speech estimate is further enhanced by applying NMF reconstruction, where this stage approximates the masked speech as a linear combination of speech model vectors. The masking stage removes the need for a noise model during reconstruction. Performing NMF reconstruction after masking provides quality improvements, but using reconstruction to approximate the masked speech is a limiting factor since the mask may contain errors that degrade perceptual quality.

Deep neural networks (DNNs) have been able to map noisy speech features to various targets such as: ideal binary masks (IBMs), ideal ratio masks (IRMs), cochleagrams, and spectrograms [9, 10, 11, 12, 13]. This has inspired us to train a DNN with a different target mapping and now we propose to use a DNN for estimating NMF activation matrices from clean speech. We will use two stages of DNNs to separate speech from background noise. In the first stage, a DNN will be trained using the IRM as target, where the ratio mask will be applied to the mixture to get a speech estimate. Features will be extracted from this speech estimate and then the second DNN will learn a mapping from the ratio-masked speech features to NMF activation matrices of clean speech. The product between the trained speech model and the estimated activation matrix will provide an estimate of clean speech in the T-F domain. The initial DNN is part of the feature extraction stage for the second DNN.

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048), and the Ohio Supercomputer Center.

This paper is organized as follows. The next section relates our work to previous studies. Section 3 describes the details of our proposed approach. Experimental results and system comparisons are presented in Section 4. Section 5 concludes the discussion of the proposed system.

2. RELATION TO PRIOR WORK

Two-stage approaches for improving speech quality are presented in our previous work [7, 8]. In [8], a ratio mask that is constructed from a binary mask is used to separate speech from background noise; then NMF is used for reconstruction. Likewise, in [7], a soft mask is used for separation followed by NMF reconstruction. Our proposed approach differs from these in the use of a DNN to estimate the ideal ratio mask (i.e. not the IBM), and a sliding window to augment our features and ground-truth training labels. We also use a second DNN to estimate the NMF activations of clean speech, not NMF reconstruction. The work in [14] also estimates NMF activations, but it estimates activations that are combined with a basis matrix to approximate the IRM. In addition, they estimate the activations of clean speech, and they do not use an initial masking stage.

3. ALGORITHM DESCRIPTION

3.1. Feature extraction

The first phase of feature extraction uses a DNN to estimate the IRM. The DNN is trained from the following complimentary set of features that are extracted from the 64-channel gammatone filter response of noisy speech: amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), and Mel-frequency cepstral coefficients (MFCC), as well as their deltas [15]. Unlike [15], the features are extracted and a single DNN is trained from the noisy speech and not separately for each frequency subband. The DNN is trained to estimate the ideal ratio mask, which is defined as:

$$IRM(t,f) = \frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)} \quad (1)$$

where $IRM(t,f)$ denotes the gain at time frame t and frequency channel f. $S^2(t,f)$ and $N^2(t,f)$ represent the clean speech and noise energy, respectively. The IRM has been shown to be the proper training target for DNN mapping, as it has outperformed other targets such as IBMs, target binary masks (TBMs), cochleagrams, and spectrograms in terms of speech quality and intelligibility [11].

A context window is used for the features and training targets of the DNN, meaning that for each time frame adjacent frames (before and after) are reshaped into a feature vector for that time frame. In other words, the DNN maps a set of frames of features to a set of frames of ground-truth IRM labels for each time frame. A context window is used for the features since useful information is carried across frames, while it is used for the labels since it has been shown to improve voice activity detection [16]. The DNN output is appropriately unwrapped and averaged to produce an estimate of the IRM, which is applied to the cochleagram of the noisy speech to produce a speech estimate. The DNN for this phase is referred to as IRM-DNN.

The second phase of feature extraction computes the log-magnitude spectrogram of the ratio-masked speech and then uses a sliding window to concatenate adjacent frames into a single feature vector for each time frame. These features are normalized to have zero mean and unit variance and are then used to train the second DNN.

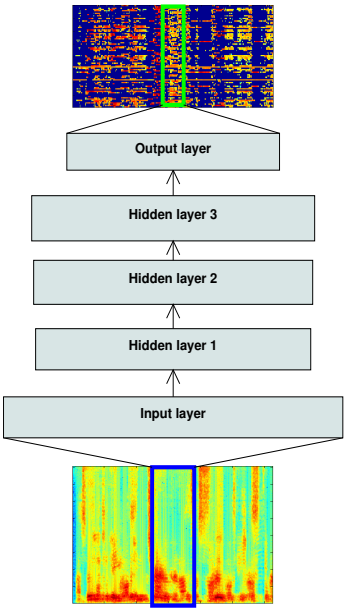


Fig. 1. Structure of DNN that maps a sliding window of log-magnitude spectrogram features from ratio-masked speech to a single frame of clean speech NMF activations.

3.2. DNN for NMF activation matrix estimation

A depiction of the second DNN is shown in Fig. 1, where the DNN consists of three hidden layers each includes 1024 hidden units. The input for each training sample is the log-magnitude spectrogram in a window of 5 frames, equating to 1285 input units. The output is the NMF activation weights in the current frame, corresponding to 80 output units, since 80 basis vectors are used for the NMF basis matrix. This DNN is referred to as NMF-DNN.

The NMF activation weights are determined from clean speech. More specifically, a NMF basis matrix is trained from a set of clean spectrograms [17]. NMF activation matrices are then computed from a second set of clean speech training data, where these activations linearly combine the elements

of the NMF basis matrix to approximate their spectrograms. This second set of clean training data is combined with various noises at different SNRs and processed through the IRM-DNN to produce a ratio mask that is subsequently applied to the mixture to generate ratio-masked speech. Log-magnitude spectrogram features are extracted from these signals and are used to train the second DNN to estimate the clean NMF activations. The clean NMF activations are also slightly modified before training, so that only the activations with values above the average activation amount for each time frame are used. This is done since the activation vector contains many small values that do not contribute much in listening quality to the result.

Once the clean NMF activations are estimated, the product between the NMF basis matrix and the estimated activation matrix is taken to produce the estimated log-magnitude spectrogram of the clean speech signal.

4. EXPERIMENTS

4.1. Experimental settings

The performance of our system is evaluated by constructing training, development, and testing data from the IEEE male speech corpus [18], after each signal is downsampled to 16 kHz. Datasets are developed for both DNNs that are used. The IRM-DNN is trained by combining 250 utterances with random cuts from babble, factory, speech-shaped noise (SSN), and military vehicle noise at -6, -3, and 0 dB, resulting in 3000 training utterances. A development set of 30 sentences mixed with each combination of noise and SNR is used to fine-tune parameters for the IRM-DNN. The NMF-DNN is trained by combining a different set of 250 utterances with random cuts of the noises at each SNR. These 3000 examples are each processed through the trained IRM-DNN, where log-magnitude spectrogram features are subsequently extracted from the ratio-mask speech. The spectrograms are computed using a window length of 32 ms, a 512 length FFT, with 75% overlap between adjacent segments. A Hann window is also used. The same development set used to train IRM-DNN is also used to train the NMF-DNN. The NMF basis matrix is trained from the concatenation of magnitude spectrograms from 10 clean speech utterances, using the above spectrogram parameters and a context window that spans 5 frames. The complete system is tested with a unique set of 720 noisy speech mixtures (60 clean utterances x 4 noises x 3 SNRs).

4.2. Results

Objective metrics PESQ [19] and the short-time objective intelligibility (STOI) [20] are used to evaluate the speech quality and intelligibility, respectively, of our system since they have been shown to correlate well with subjective quality and intelligibility evaluations from human subjects.

Table 1. Average PESQ and STOI scores for the different systems at each SNR.

	PESQ			STOI		
	-6 dB	-3 dB	0 dB	-6 dB	-3 dB	0 dB
Noisy Speech	1.650	1.816	1.990	0.584	0.641	0.701
SM/NMF	2.037	2.119	2.188	0.643	0.689	0.724
IRM/NMF	2.055	2.130	2.195	0.656	0.696	0.727
N-FHMM	1.841	1.976	2.141	0.580	0.632	0.695
NMF	1.939	2.110	2.285	0.632	0.694	0.754
Proposed	2.370	2.570	2.736	0.775	0.820	0.851
Prop. w/o IRM	2.394	2.548	2.675	0.782	0.824	0.854

Table 2. Average PESQ scores for the different systems at each noise type.

	PESQ			
	Babble	Factory	SSN	Vehicle
Noisy Speech	1.728	1.631	1.669	2.247
SM/NMF	2.081	2.063	2.115	2.199
IRM/NMF	2.085	2.106	2.107	2.208
N-FHMM	1.823	1.803	1.880	2.438
NMF	1.961	1.872	1.951	2.661
Proposed	2.492	2.496	2.420	2.827
Prop. w/o IRM	2.503	2.508	2.436	2.710

We compare our approach to four separate systems [7, 17, 21], two NMF approaches and two systems that incorporate masking and NMF reconstruction. A supervised NMF approach from [17] uses trained speech and noise models to approximate noisy speech. The speech model matches the NMF basis matrix that we use, while the noise model is trained from the concatenated spectrograms of all the noise signals. The work in [21] uses a semi-supervised nonnegative factorial hidden Markov model (N-FHMM) that incorporates a non-negative hidden Markov model (N-HMM) as the speech model, while a noise model is learned during testing. N-HMM uses several small dictionaries, each of which represent a particular phoneme, and a HMM to model transitions between different phonemes. The N-HMM is trained from the 10 clean speech utterances that are used for the NMF basis matrix. Since our goal is to show that using a DNN to determine activation weights is better than using NMF reconstruction, we compare our system to [7] (i.e. SM/NMF) and a system that uses an estimated IRM to separate speech from noise, followed by NMF reconstruction (i.e. IRM/NMF). Both of these approaches use DNNs to generate a mask, but [7] uses a soft mask in its first stage, where the IBM is used as a ground-truth label. Context windows are used to modify the spectrograms for each of the models.

Table 1 shows the average PESQ and STOI performance for each system at each SNR. Note that at -6 dB SNR conditions, the two stage approaches offer much quality improvement over the NMF based approaches, indicating that a masking stage that removes noise is important. It also indicates

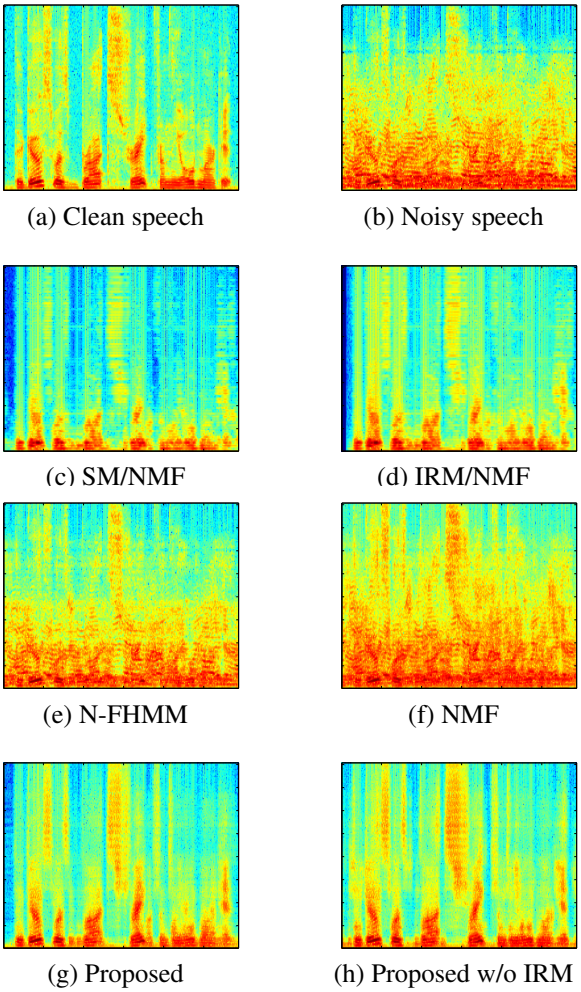
Table 3. Average STOI scores for the different systems at each noise type.

	STOI			
	Babble	Factory	SSN	Vehicle
Noisy Speech	0.570	0.588	0.605	0.805
SM/NMF	0.667	0.642	0.686	0.746
IRM/NMF	0.672	0.658	0.692	0.749
N-FHMM	0.576	0.583	0.605	0.780
NMF	0.647	0.635	0.646	0.844
Proposed	0.808	0.789	0.797	0.866
Prop. w/o IRM	0.817	0.791	0.808	0.864

that at very low SNRs, portions of the speech activations from NMF and N-FHMM approximate some of the noise components. Our proposed approach offers significant PESQ and STOI improvements over the four comparison systems at each SNR, indicating that a masking DNN and DNN for estimating clean activation matrices are beneficial. The NMF-DNN offers improvements over NMF reconstruction because some of the mistakes that are made during the mask estimation stage can be corrected in the second-stage DNN and masked speech energy can be restored. IRM/NMF also offers slight improvements over SM/NMF because the estimated IRM outperforms soft masking, which matches results from [11] and justifies its use as the first phase of feature extraction for our proposed algorithm. Fig. 2 shows spectrogram results for the different systems at -3 dB with babble noise. Notice that portions of the speech are removed in the IRM/NMF and SM/NMF approaches, but some of the speech energy is restored in the proposed signal.

Table 2 shows the PESQ performance of the systems averaged over the different SNR levels, for each noise type. From these results we see that NMF and N-FHMM struggle with speech-like and non-stationary noises such as babble and factory. The proposed method substantially outperforms the IRM/NMF method at each noise type. All systems show much improvement when the mixture contains military vehicle noise. Similar performance results are seen in Table 3, which shows the average objective intelligibility of the systems at each noise type.

As a final comparison, a DNN is trained that maps the log-magnitude spectrograms from noisy speech utterances to clean speech activations. This DNN is trained with the same utterances used to train our proposed system, however, each utterance is not processed with the IRM-DNN. The different results for this system are shown in Tables 1-3 as 'Prop. w/o IRM'. Table 1 shows that the STOI performance for this system and the proposed are approximately identical. In terms of PESQ, the proposed system offers slight improvements over 'Prop w/o IRM' at -3 dB and 0 dB, but performs slightly worse at -6 dB. The performance by noise type (i.e. Table 2) shows that the speech quality of the signals is approximately equal at each noise type, except military vehicle noise where



(a) Clean speech (b) Noisy speech (c) SM/NMF (d) IRM/NMF (e) N-FHMM (f) NMF (g) Proposed (h) Proposed w/o IRM

Fig. 2. Example spectrograms of different signals at -3 dB using babble noise.

the proposed approach performs better. Table 3 shows that the STOI scores for the two approaches are approximately the same across the different noise types. Fig. 2 shows that the proposed method that does not use ratio masking may overemphasize some speech components as compared to the clean speech, which is indicated by observing some of the high-frequency components of the unvoiced frames.

5. CONCLUSION

We have proposed to use DNNs to estimate the NMF activation matrices of clean speech. The first DNN estimates the ideal ratio mask and is part of the feature extraction stage for the second DNN. The second DNN estimates the NMF activation weights from ratio-masked speech. The algorithm improves objective speech quality and intelligibility at various noisy conditions. The results show that this approach outperforms similar techniques.

6. REFERENCES

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

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

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

[4] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 15, pp. 1–12, 2007.

[5] 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.

[6] C. Févotte, N. Bertin, and J-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.

[7] D. S. Williamson, Y. Wang, and D. L. Wang, "A two-stage approach for improving the perceptual quality of separated speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014, pp. 7084–7088.

[8] D. S. Williamson, Y. Wang, and D. L. Wang, "Reconstruction techniques for improving the perceptual quality of binary masked speech," *The Journal of the Acoustical Society of America*, vol. 136, pp. 892–902, 2014.

[9] 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.

[10] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7092–7096.

[11] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, pp. 1849–1858, 2014.

[12] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014, pp. 4661–4665.

[13] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, pp. 65–68, 2014.

[14] Y. Wang and D. L. Wang, "A structure-preserving training target for supervised speech separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014, pp. 6148–6152.

[15] 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.

[16] X.-L. Zhang and D. L. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. of INTERSPEECH*, 2014, pp. 1534–1538.

[17] J. Eggert and E. Korner, "Sparse coding and NMF," in *Proc. Neural Networks*, 2004, vol. 4, pp. 2529–2533.

[18] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[19] 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.

[20] 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.

[21] 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.