Estimating nonnegative matrix model activations with deep neural networks to increase perceptual speech quality

Donald S. Williamson[a] and Yuxuan Wang
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

DeLiang Wang
Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA

(Received 12 February 2015; revised 3 June 2015; accepted 1 August 2015; published online 10 September 2015)

As a means of speech separation, time-frequency masking applies a gain function to the time-frequency representation of noisy speech. On the other hand, nonnegative matrix factorization (NMF) addresses separation by linearly combining basis vectors from speech and noise models to approximate noisy speech. This paper presents an approach for improving the perceptual quality of speech separated from background noise at low signal-to-noise ratios. An ideal ratio mask is estimated, which separates speech from noise with reasonable sound quality. A deep neural network then approximates clean speech by estimating activation weights from the ratio-masked speech, where the weights linearly combine elements from a NMF speech model. Systematic comparisons using objective metrics, including the perceptual evaluation of speech quality, show that the proposed algorithm achieves higher speech quality than related masking and NMF methods. In addition, a listening test was performed and its results show that the output of the proposed algorithm is preferred over the comparison systems in terms of speech quality.
 [http://dx.doi.org/10.1121/1.4928612]

[MAH] Pages: 1399–1407

I. INTRODUCTION

The performance of many speech processing applications and devices is impaired by the presence of background noise. The accuracy of automatic speech recognition and speaker identification systems is degraded when background noise interferes with speech. Hearing aids that rely on signal amplification amplify both speech and noise. Many computational techniques address monaural speech separation in noisy environments, but only a few studies based on time-frequency (T-F) masking show improved speech intelligibility (Kim *et al.*, 2009; Healy *et al.*, 2013). In these studies, speech quality remains an issue.

Nonnegative matrix factorization (NMF) has been extensively used for separating speech from background noise. NMF uses the product of a basis matrix and activation matrix to approximate a signal, where the basis matrix provides spectral structure and the activation matrix linearly combines the basis matrix elements (Lee and Seung, 1999; Seung and Lee, 2001). The main goals of NMF are to train an appropriate basis matrix that provides a generalized spectral representation of speech and noise, and to determine an activation matrix that 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 (Virtanen, 2007; Smaragdis, 2007; Wilson *et al.*, 2008; Févotte *et al.*, 2009). The basis matrix is obtained by concatenating a speech model and a noise model. The 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 models to generate an approximation of the speech and noise components of the mixture. For instance, the first portion of the activation matrix is combined with the speech basis matrix to approximate the speech portion of the mixture, whereas the second portion of the activation matrix and the noise model estimate the noise. The speech and noise estimates are then combined to produce a Wiener-like mask applied to the mixture to extract the speech. Supervised NMF has been shown to improve the objective quality of separated speech, however, it has not been shown to improve the intelligibility of extracted speech.

A recent modification to supervised NMF is the nonnegative factorial hidden Markov model (N-FHMM) (Mysore and Smaragdis, 2011). 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 (Mysore *et al.*, 2010). A single dictionary is selected to approximate the speech in each time frame. N-FHMM produces a Wiener mask that is used to separate the speech from the noise. The challenge with N-FHMM is how to ensure that the appropriate dictionary is used in each frame, which may not always occur.

In Williamson *et al.* (2014a,b), we have shown that combining a T-F masking approach with NMF reconstruction produces higher quality speech than supervised NMF and masking alone, and other two-stage methods. A deep neural

[a]Electronic mail: williado@cse.ohio-state.edu

network (DNN) is used to estimate a T-F mask (binary and soft) that, when applied to the noisy mixture, produces a speech estimate. This speech estimate is further enhanced by applying NMF reconstruction, which approximates the masked speech by linearly combining elements from a speech model. Importantly, the masking stage removes the need for a noise model during reconstruction. However, using reconstruction to approximate the masked speech is a limiting factor since the mask may contain errors that degrade perceptual quality.

DNNs have been used to estimate various targets such as ideal binary masks (IBMs), ideal ratio masks (IRMs), and spectrograms (Wang and Wang, 2013; Narayanan and Wang, 2013; Xu *et al.,* 2014; Han *et al.,* 2014). In this study, 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, where the initial DNN is part of the feature extraction stage for the second DNN. In the feature extraction stage, a DNN will be trained to approximate the IRM, and the ratio mask will be applied to the mixture to get a speech estimate. Temporal correlations will be accounted for by using a sliding window to augment the features and training labels for the DNN. Features will be then extracted from this speech estimate and 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. Noticeable artifacts are produced when estimated IBMs or IRMs are used as training targets for DNNs, however, they effectively suppress background noise. On the other hand, NMF (and related techniques) provide good approximations when they are used to estimate clean speech. Using DNNs to estimate the clean model activations, after denoising with an estimated IRM, should produce estimates that have fewer unwanted artifacts and are closer to clean speech.

Two-stage methods for improving speech quality are presented in our previous work (Williamson *et al.,* 2014a,b). In Williamson *et al.* (2014a), a soft mask is used to separate speech from background noise; then NMF is used for reconstruction. The soft mask is estimated using a DNN, where the IBM is used as the training target. In Williamson *et al.* (2014b), a ratio mask constructed from a binary mask is used for separation, followed by NMF reconstruction. Our proposed approach significantly differs from these earlier studies in several ways. The first is the use of a DNN to estimate the ideal ratio mask, i.e., not the IBM. Second, the present study uses a sliding window to augment features and training labels. Third, we use a second DNN to estimate the NMF activations of clean speech, not NMF reconstruction.

A number of evaluations are performed to assess the overall quality of separated speech at different signal-to-noise ratios and with different interferences. Our proposed approach and relevant comparison systems are evaluated with the perceptual evaluation of speech quality (PESQ) measure. In addition, a listening study is conducted where individuals with normal hearing compare pairs of signals and select the preferred signal in terms of quality. The results show that the proposed algorithm produces higher speech quality and is preferred over comparison systems.

The rest of the paper is organized as follows. The proposed approach is described in Sec. II. Section III evaluates and compares the proposed system with related approaches using objective metrics. The human listening study is described in Sec. IV. Finally, concluding remarks are given in Sec. V.

II. ALGORITHM DESCRIPTION

A diagram of the proposed approach is given in Fig. 1. In the feature extraction stage, an ideal ratio mask is estimated from the noisy speech mixture using a deep neural network. The estimated IRM is applied to the cochleagram of the mixture to produce a speech estimate. The resulting separated speech is augmented by incorporating temporal continuity between successive time frames. A second deep neural network, using the speech separated by ratio masking as input, estimates the NMF activation matrix of clean speech. The product between the speech model and the estimated activation matrix gives an approximation of clean speech STFT (short-time Fourier transform) magnitude response. Finally, the estimated STFT magnitude response is combined with the STFT phase response of noisy speech to produce the final estimate of the speech signal using overlap-and-add synthesis. Sections II A and II B describe these steps in more detail.

A. Feature extraction

The first phase of feature extraction uses a DNN to estimate the IRM. The DNN is trained from the following complementary set of features that are extracted from the gammatone filter responses of noisy speech: amplitude modulation spectrogram, relative spectral transform and perceptual linear prediction, and Mel-frequency cepstral coefficients, as well as their deltas (Wang *et al.,* 2013). The DNN is trained to estimate the ideal ratio mask, which is defined as

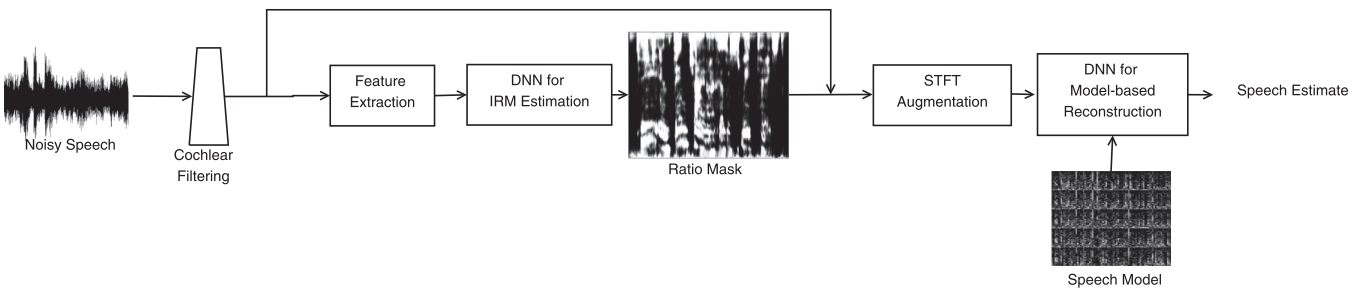

FIG. 1. Block diagram of the proposed approach.

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

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 gain at each T-F unit may be interpreted as the proportion of energy that is attributed to speech, so when the IRM is applied to noisy speech, each T-F unit retains the correct amount of speech energy. The IRM has been shown to be more effective for DNN mapping than other training targets such as the IBM, the target binary mask, cochleagram, and spectrogram in terms of combined speech quality and intelligibility (Wang *et al.*, 2014).

We employ a context window 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 IRM labels around each time frame. A context window is used because useful speech information is carried across time frames. In addition its utility has been shown in a recent study on voice activity detection (Zhang and Wang, 2014). 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 the IRM-DNN.

Specifically, the IRM-DNN consists of four layers (three hidden, one output), where the hidden layers each have 1024 nodes. The hidden nodes use rectified linear activation functions (Nair and Hinton, 2010), while the output nodes use a sigmoidal activation function. The input and hidden layers of the DNN are trained with a dropout rate of 0.2. The weights are randomly initialized to values between -1 and 1, and they are updated using mini-batch stochastic gradient descent with ADAGRAD (Duchi *et al.*, 2010) and momentum. The mini-batch size is set to 512 and the scaling factor for ADAGRAD is 0.005. The momentum is set to 0.5 for the first 5 epochs and to 0.9 thereafter. The DNN is trained to minimize the mean-square error over 150 epochs. A development set was used to determine these parameter values, where the set of values that minimized the cost function is used.

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.

B. DNN for NMF activation matrix estimation

A depiction of the second DNN is shown in Fig. 2. The input to this DNN is the log-magnitude spectrogram of ratio-masked speech, which is computed as follows:

$$\hat{\boldsymbol{S}}_1(t,f) = \boldsymbol{eIRM}(t,f) \odot \boldsymbol{Y}(t,f), \quad (2)$$

where $\odot$ denotes the Hadamard product (element-wise multiplication), $\boldsymbol{eIRM}$ is the estimated IRM, $\boldsymbol{Y}$ is the spectrogram of the noisy speech, and $\hat{\boldsymbol{S}}_1$ is the ratio-masked speech

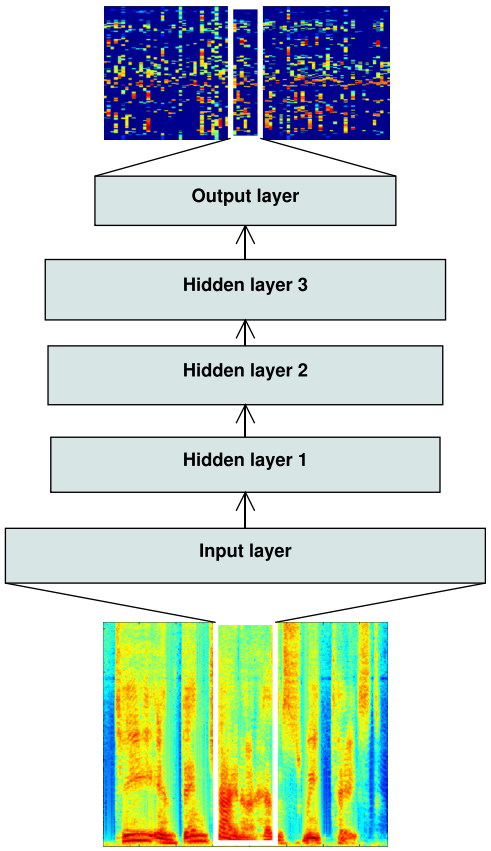

FIG. 2. (Color online) Structure of DNN that maps a sliding window of log-magnitude spectrogram features from ratio-masked speech to a single frame of clean speech NMF activations.

estimate. The subscript 1 indicates that it is the spectrogram estimate after the first DNN. The training target for this DNN is the NMF activation weights in the current frame. This DNN is referred to as NMF-DNN. The parameters for the NMF-DNN match those of the IRM-DNN.

For training, a NMF basis matrix, $\boldsymbol{W}_{tr}$, is iteratively trained from a set of clean spectrograms, $\boldsymbol{D}$, using (Eggert and Korner, 2004)

$$\begin{aligned}
&\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_N];\\
&\boldsymbol{H}_{tr} \leftarrow \boldsymbol{H}_{tr} \odot \frac{\boldsymbol{W}_{tr}^T\left(\frac{\boldsymbol{D}}{\boldsymbol{W}_{tr}\boldsymbol{H}_{tr}}\right)}{\boldsymbol{W}_{tr}^T\mathbf{1}_H + \lambda}\\
&\boldsymbol{W}_{tr} \leftarrow \boldsymbol{W}_{tr} \odot \frac{\left(\frac{\boldsymbol{D}}{\boldsymbol{W}_{tr}\boldsymbol{H}_{tr}}\right)\boldsymbol{H}_{tr}^T + \mathbf{1}_W\left(\mathbf{1}_H\boldsymbol{H}_{tr}^T \odot \boldsymbol{W}_{tr}\right)}{\mathbf{1}_H\boldsymbol{H}_{tr}^T + \mathbf{1}_W\left(\left(\frac{\boldsymbol{D}}{\boldsymbol{W}_{tr}\boldsymbol{H}_{tr}}\right)\boldsymbol{H}_{tr}^T \odot \boldsymbol{W}_{tr}\right)},
\end{aligned} \quad (3)$$

where $\mathbf{1}_H$ is an all-one matrix with the same dimensions as $\boldsymbol{D}$, and $\mathbf{1}_W$ is an all-one square matrix. $\boldsymbol{D}$ is generated by concatenating the clean spectrograms from a set of N clean speech signals (i.e., $\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_N$). λ is a parameter that controls sparseness and it is set to 0.1 (see Williamson *et al.*, 2014b). All divisions in Eq. (3) represent element-wise division. $\boldsymbol{H}_{tr}$ is the trained activation matrix that linearly combines the elements of $\boldsymbol{W}_{tr}$ so that their product approximates $\boldsymbol{D}$ [Eq. (4)].

As evaluated in Williamson *et al.* (2014b), the generalized KL-divergence outperforms other cost functions in terms of PESQ performance, so this is what we use in this current study. Once the NMF basis matrix is determined, the trained activation matrix is discarded:

$$\boldsymbol{D} \approx \boldsymbol{W}_{tr}\boldsymbol{H}_{tr}. \tag{4}$$

NMF activation matrices, $\boldsymbol{H}_{S_2}$, from clean speech spectrograms, $\boldsymbol{S}_2$, are then computed separately for each signal in a second set of clean speech training data, where these activations linearly combine the column vectors of $\boldsymbol{W}_{tr}$ to approximate the clean speech spectrograms (i.e., $\boldsymbol{S}_2 \approx \boldsymbol{W}_{tr}\boldsymbol{H}_{S_2}$). The clean NMF activations are iteratively computed using Eq. (5). The clean activations are modified so that only the activations with values above the average activation amount in each time frame are retained, and activations below the average are set to zero. This is done since the activation vector contains many small values that do not contribute much to listening quality of the result:

$$\boldsymbol{H}_{S_2} \leftarrow \boldsymbol{H}_{S_2} \odot \frac{\boldsymbol{W}_{tr}^T\left(\frac{\boldsymbol{S}_2}{\boldsymbol{W}_{tr}\boldsymbol{H}_{S_2}}\right)}{\boldsymbol{W}_{tr}^T\boldsymbol{1}_H + \lambda}. \tag{5}$$

Each signal in this 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 the ratio-masked signals and used to train the second DNN to estimate the clean NMF activations. This DNN minimizes the mean-square error between the clean NMF activations and its estimated versions.

Once the clean NMF activations are estimated using the second DNN, the product between the NMF basis matrix and the estimated activation matrix, $\hat{\boldsymbol{H}}_{S_2}$, is taken to produce the estimated spectrogram of the clean speech signal using Eq. (6),

$$\hat{\boldsymbol{S}}_2 = \boldsymbol{W}_{tr}\hat{\boldsymbol{H}}_{S_2}. \tag{6}$$

$\hat{\boldsymbol{S}}_2$ is the spectrogram estimate after the second DNN. This estimate is combined with the noisy phase, and then overlap-and-add synthesis is used to produce the final time-domain estimate.

III. EXPERIMENT I: OBJECTIVE EVALUATION AND COMPARISON

A. Experimental setup

Our system is developed and evaluated by constructing training, development, and testing sets from the IEEE male speech corpus (IEEE, 1969), where datasets are developed for each DNN. The signals are downsampled to 16 kHz prior to processing. 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 (250 utterances × 4 noises × 3 SNRs). The features into IRM-DNN are extracted from the 64-channel gammatone filter response of noisy speech. Unlike Wang *et al.* (2013), the features are extracted at the frame level (i.e., not per frequency channel) and a single DNN is trained from the noisy speech. 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 the log-magnitude spectrogram is then computed from the ratio-mask speech. The spectrograms are computed using a window length of 32 ms, a 512 point FFT, and 75% overlap between adjacent segments. A window of five frames is used for the context window, resulting in 1285 [i.e., $(512/2+1)5$] input units into NMF-DNN. A Hann window is used. NMF-DNN consists of three hidden layers each including 1024 hidden units. The same development set used to determine parameter values for the IRM-DNN is also used to determine parameter values for the NMF-DNN. The NMF basis matrix is trained from the concatenation of magnitude spectrograms from ten clean speech utterances, using the above spectrogram parameters and a context window that spans five frames. The NMF basis matrix consists of 80 basis vectors. The complete system is tested with a unique set of 720 noisy speech mixtures (60 clean utterances × 4 noises × 3 SNRs), where the random cuts of noise used in the testing mixtures do not overlap with the random cuts of noise used for the training and development mixtures.

PESQ (ITU-R, 2001) and the short-time objective intelligibility (STOI) (Taal *et al.*, 2011) are used to evaluate the speech quality and intelligibility, respectively. These objective metrics have been shown to correlate well with perceptual quality and intelligibility evaluations from human subjects.

We compare our approach to four related systems, where two are NMF approaches and two systems incorporate masking and NMF reconstruction. The supervised NMF approach from Eggert and Korner (2004) uses trained speech and noise models to approximate noisy speech. The speech model matches the NMF basis matrix we use, while the noise model is trained from the concatenated spectrograms of all the noise signals. The work in Mysore and Smaragdis (2011) uses a semi-supervised nonnegative factorial hidden Markov model (N-FHMM) to separate speech from background noise. It uses a non-negative hidden Markov model (N-HMM) as the speech model, and a noise model is learned during testing. N-HMM uses several small dictionaries, each of which represents a particular phoneme, and an HMM to model transitions between different phonemes. The N-HMM is trained from the same ten clean speech utterances used for training the NMF basis matrix. The noise model is represented with a single large dictionary. Since our goal is to demonstrate that using a DNN to determine activation weights is better than using NMF reconstruction, we compare our system to Williamson *et al.* (2014a) (i.e., SM/NMF with SM standing for soft masking) and a system that uses an estimated IRM and NMF reconstruction to separate speech from noise (i.e., IRM/NMF). Both of these approaches use DNNs to generate a mask, but Williamson *et al.* (2014a) uses a soft mask in its first stage, where the

IBM is used as a ground-truth label during training. Context windows are also used to modify the features, DNN outputs, and NMF basis matrix for SM/NMF and IRM/NMF, and they are used to augment the NMF and N-FHMM models.

B. Results and discussions

Table I shows the average PESQ and STOI scores for each system at each SNR. Note that at −6 dB, the approaches that perform masking and reconstruction offer quality improvement over the NMF based approaches (supervised and semi-supervised), indicating that a masking stage removing noise is important. Our proposed approach offers significant PESQ and STOI improvements over the four comparison systems at every SNR. Informal listening reveals that NMF and N-FHMM outputs contain residual noise and speech distortions. This is likely due to inaccuracies in determining the speech and noise model activations that combine to approximate noisy speech. In other words, the product of the NMF basis matrix and estimated speech model activations does not effectively approximate the clean speech in the noisy speech signal. These approaches often struggle when there is a high amount of noise, since they operate independently on mixtures without prior training. The NMF-DNN outperforms NMF reconstruction because some of the mistakes during the mask estimation stage can be corrected by the second-stage DNN. Estimating clean speech using the NMF-DNN, as opposed to estimating masked-speech with NMF reconstruction, helps to remove artifacts due to IRM estimation errors. IRM/NMF offers slight improvements over SM/NMF because the estimated IRM outperforms soft masking, consistent with the observation from Wang *et al*. (2014) and justifying the use of the IRM as the first phase of feature extraction for our proposed algorithm. The IRM/NMF outperforms SM/NMF likely because predicting ratio targets is less sensitive to estimation errors than predicting binary targets, since errors in binary decisions may be more costly to speech quality. Figure 3 illustrates 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 that speech energy is restored in the proposed signal. The spectrograms from supervised NMF and semi-supervised N-FHMM show that at this low SNR, portions of the speech activations approximate noise components, which is indicated by the prevalence of remaining noise.

TABLE I. Average PESQ and STOI scores for different systems at each SNR. Bold indicates best result.

	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**

Table II shows the PESQ performance of the systems averaged over the different noise types. From these results we see that NMF and N-FHMM improve objective speech quality a little bit, and SM/NMF and IRM/NMF improve a little more. The proposed method substantially outperforms both kinds of method for each noise type. When the mixture contains military vehicle noise, the two masking based methods do not lead to improvement over unprocessed noisy speech. Similar results are seen in Table III, which shows the average objective intelligibility of the systems for the different noise types.

IV. EXPERIMENT II: HUMAN SUBJECT TESTING

Although our proposed approach outperforms comparison systems in quality and intelligibility based on objective metrics, the true indication of quality improvement must come from human listeners. Thus, we conduct a listening study to determine if human subjects prefer the quality of our proposed approach over others. More specifically, we compare our proposed approach described in Sec. II to speech separated by an estimated IRM and speech separated using N-FHMM (Mysore and Smaragdis, 2011). We use N-FHMM over NMF since the training and testing methodologies of the former are consistent with the proposed algorithm, where only a speech model is trained (i.e., it has no access to the noise signal that is present in a mixture). Comparing the proposed algorithm to speech separated by an estimated IRM allows us to determine if using a deep neural network to predict clean NMF activations offers quality improvements over ratio masking alone. Likewise, comparing against N-FHMM addresses the question of whether our algorithm improves perceptual quality over a model-based approach that uses traditional methods to determine activations.

A. Experimental setup

Our listening study compares pairs of signals and has the participant select the signal that they prefer in terms of quality. A preference rating is often used in speech quality evaluations, particularly for quality comparisons (Arehart *et al*., 2007; Koning *et al*., 2015). The listeners are instructed to play each signal at least once. After listening to the pair of signals, the subject is instructed to select one of three options: signal A is preferred, signal B is preferred, or neither signal is preferred over the other. The last option indicates that the qualities of the signals are approximately identical. Pairwise scoring is employed, with the score of +1 awarded to the preferred method and −1 to the other. For a similar-preference response (i.e., the third option) each method is awarded the score of 0. If the participant indicates that signal A or signal B is better, than they provide an improvement score, ranging from 0 to 4 with increments of 0.01 for the higher quality signal. A grade of 0 indicates that the quality of the two signals is identical, 1 indicates that the quality of the preferred signal is slightly better than the other signal, 2 indicates that the quality of the preferred signal is better than the other signal, while grades of 3 and 4 indicate

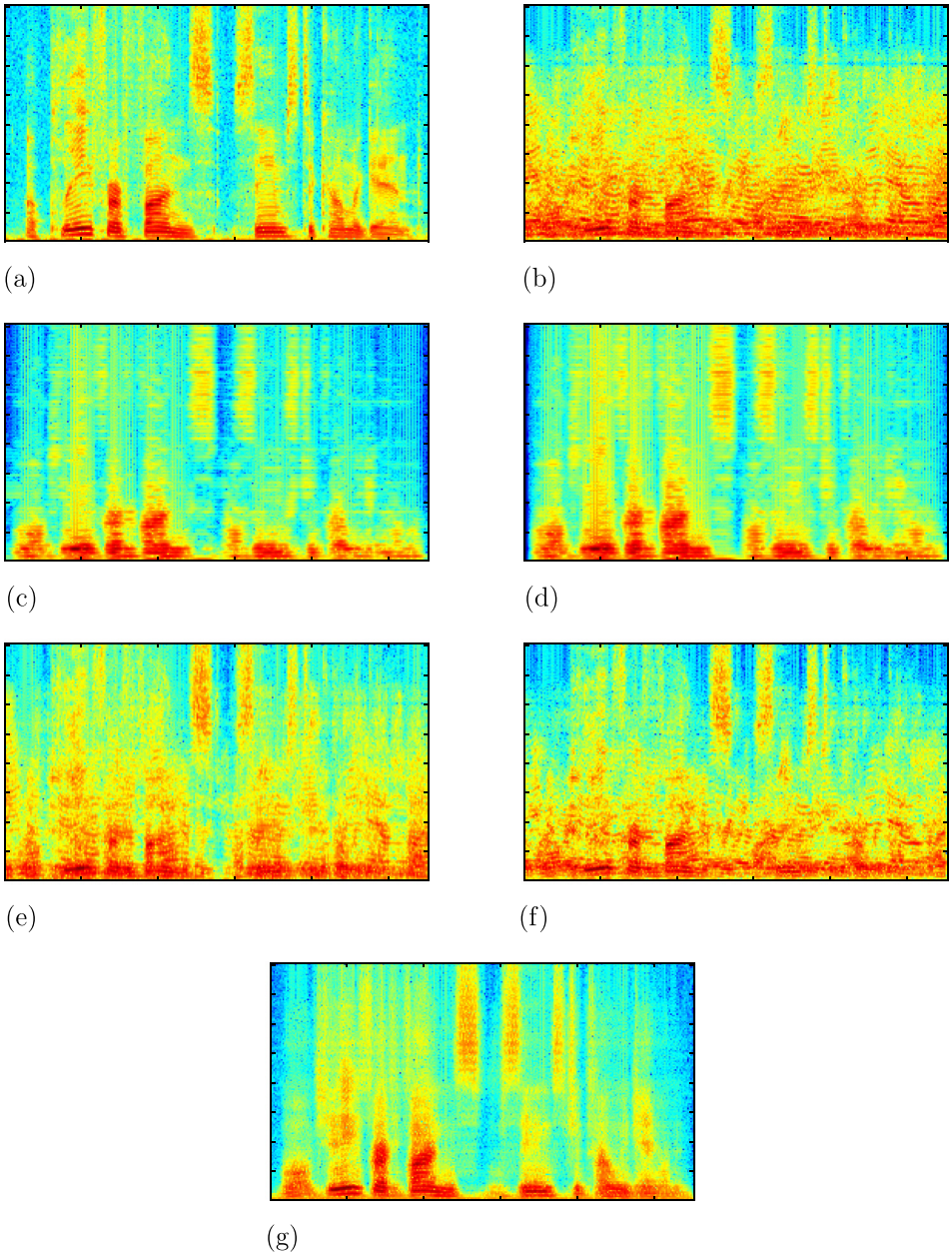

FIG. 3. (Color online) Example spectrograms of different signals at −3 dB with babble noise. (a) Clean utterance. (b) Noisy utterance. (c) Speech separated by soft masking and NMF reconstruction. (d) Speech separated by ratio masking and NMF reconstruction. (e) Speech separated by N-FHMM. (f) Speech separated using supervised NMF. (g) Speech separated by the proposed system.

that the quality of the preferred signal is largely and hugely better than the other signal, respectively (see Koning *et al.*, 2015). Figure 4 displays the graphical interface that the subject used to complete the evaluation.

The systems and signals used during this study are generated as described in Sec. III A. Only signals processed with combinations of factory, speech-shaped noise, or military vehicle noise at SNRs of −3 and 0 dB are assessed. The SNR of −6 dB is not used to ensure that the processed signals are intelligible to listeners.

The listening study consists of three phases: practice, training, and evaluation. In the practice phase the subject

TABLE II. Average PESQ scores for different systems across the noise types. Bold indicates best result.

	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**

TABLE III. Average STOI scores for different systems across the noise types. Bold indicates best result.

	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**

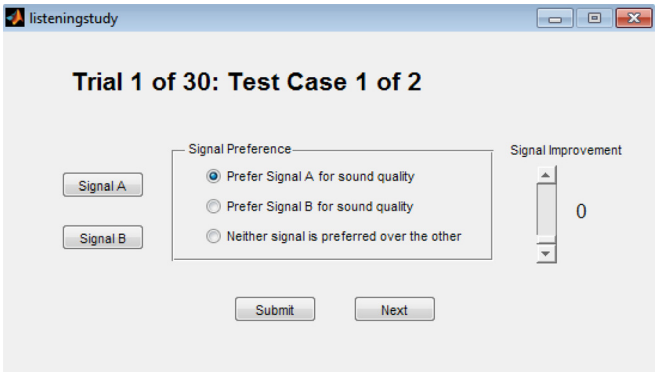

FIG. 4. (Color online) Screenshot of the graphical interface for the listening study.

listens to example sound excerpts to illustrate the whole range of qualities. Two groups of signals are presented, where each group contains a clean speech, proposed, noisy speech, N-FHMM, and estimated IRM signal. The noisy conditions (SNR and noise) for each group are different, where the first group is generated using factory noise at −3 dB SNR and the second group uses military vehicle noise at −3 dB. Ten signals in total are presented.

The training session allows the subject to become familiar with the graphical interface. The user performs three sets of evaluations, where each evaluation involves the comparisons of two pairs of signals. The following signals are compared for each evaluation set in the following order: (1) estimated IRM-processed speech to proposed-processed speech and (2) proposed-processed speech to N-FHMM processed speech. Three IEEE testing sentences that are not used during the evaluation phase or practice session are randomly chosen and used for each training set. For the three sets, the test signals are combined with SSN, factory, or military vehicle noise at a 0 dB SNR. For each pair of signals, the subject listens to each signal at least once. The test cases are presented to the listener through a MATLAB graphical user interface (GUI), where the GUI will display a single comparison at a time and a total of 6 comparisons are presented.

Upon completion of the practice and training sessions, the subject begins the formal evaluation phase of the listening study. The user performs 60 total comparisons. Five sets of each combination of SNR (−3 and 0 dB) and noise (SSN, factory, and military vehicle) are evaluated, resulting in 30 total combinations. Each combination requires two comparisons: (1) estimated IRM-processed speech to proposed-processed speech and (2) proposed-processed speech to N-FHMM processed speech. The utterance and order of presentation of the comparisons are randomly selected from the test signals, which have not been used in the practice or training sessions. The listener has no prior knowledge on the algorithm used to produce a signal. The presentation order of the different conditions is randomly generated by the GUI for each listening subject.

The signals are presented diotically over Sennheiser HD 265 Linear headphones using a personal computer, where each signal is normalized to have the same sound level. The subjects are seated in a sound proof room. The participants are instructed to play a sound as often as possible to aid in making a quality determination.

Ten subjects (six female and four male), between the ages of 21 and 33, each with self-reported normal hearing participated in the study. All the subjects are native English speakers. These were students recruited from The Ohio State University and they received a monetary incentive for participating.

B. Results and discussions

The sound quality preference scores averaged across the subjects are displayed in Fig. 5. The upper panels display the

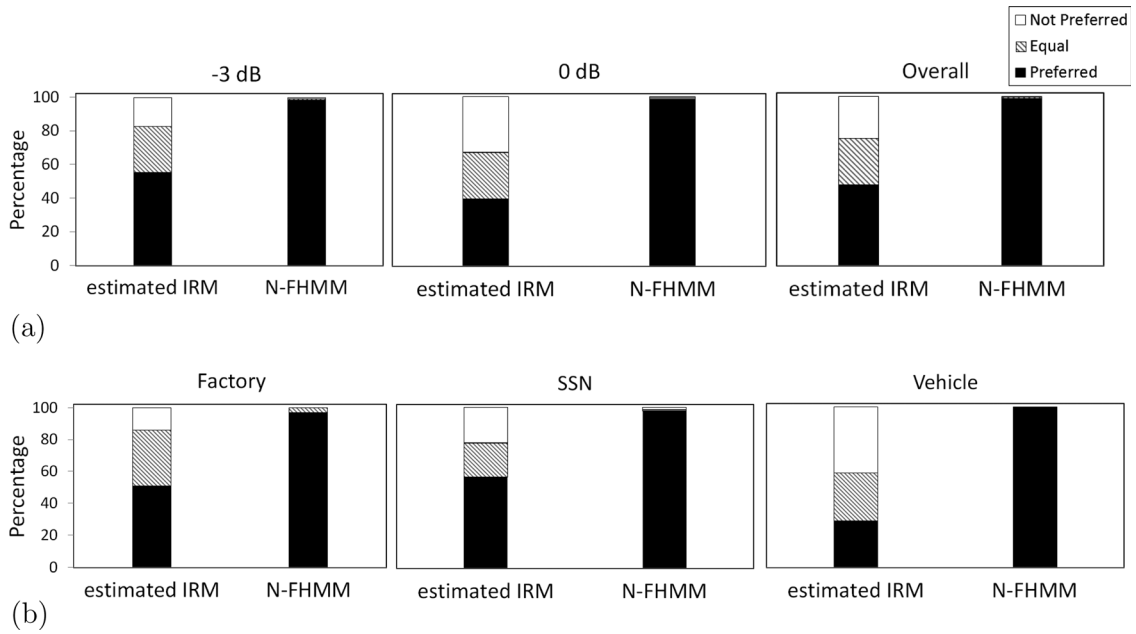

FIG. 5. Group mean sound quality preference scores for our proposed approach when evaluated against an estimated ideal ratio mask (IRM) and non-negative factorial hidden Markov model (N-FHMM) for IEEE sentences following algorithm processing.

scores averaged by signal-to-noise ratio and the overall mean sound quality preference scores. The lower panels show data for sentences in factory, speech-shaped noise, and military vehicle noise, respectively. Each panel contains the percentage that the user prefers signals produced by our proposed algorithm (i.e., preferred), the percentage that the users prefer the comparison signal (i.e., not preferred), and the percentage that the users believe that the sound quality between the two signals is equal (i.e., equal). The subjects preferences are indicated by the size of each block (i.e., preferred, not preferred, and equal) within each panel. It is clear from Fig. 5 that subjects prefer the proposed approach over comparison approaches.

As seen in the overall panel (top-right), subjects prefer the proposed approach over the estimated IRM approximately 48% of the time, they prefer the estimated IRM 25%, and they feel that the sound quality of the estimated IRM and the proposed approach are equal roughly at a rate of 27%. Thus, the sound quality of the proposed approach is equal to or better than the estimated IRM 75% of the time, indicating that the DNN-NMF helps sound quality. When compared to N-FHMM, participants almost always prefer our method (i.e., in 98% of the time).

The upper panel also displays the sound quality preference percentages for signals with −3 and 0 dB SNR. The preferred score is approximately 98% for our algorithm when compared to N-FHMM at −3 and 0 dB separately. At −3 and 0 dB, the preferred scores are around 55% and 40%, respectively, as compared to the estimated IRM, where the qualities of the signals are equal with a rate of 27%. The decrease in quality preference when the SNR increases from −3 to 0 dB is due in large part to the IRM-DNNs ability to estimate the ideal ratio mask. At lower SNRs, the quality of ratio masking is lower due to an increase in noise energy and estimation errors (Wang *et al.,* 2014), so improvements by the DNN-NMF are more pronounced.

The preference scores for our method when compared to the estimated IRM are approximately 51%, 57%, and 29% for factory, speech shaped, and military vehicle noise, respectively, as indicated by the lower panel in Fig. 5. Participants prefer estimated IRM at 14%, 22%, and 41% rates for the corresponding noises. The average equality scores for this same comparison are 35%, 21%, and 30%, respectively. For each noise type, the preference score for our algorithm when compared against N-FHMM is approximately 98%.

A one-way analysis of variance (ANOVA) test is performed to determine whether the preference results displayed in Fig. 5 are statistically significant. The test is run for each panel displayed in the figure and the *p*-value for each category is shown in Table IV. The *p*-values for each comparison and category are zero, except for the comparison between the proposed system and estimated IRM at 0 dB SNR and military vehicle noise, indicating that all results except these two cases are statistically significant.

TABLE IV. *p*-value scores for one-way ANOVA test for preference results when the proposed system is compared to estimated IRM and N-FHMM.

	Overall	−3 dB	0 dB	Factory	SSN	Vehicle
Estimated IRM	0.000	0.000	0.137	0.000	0.000	0.043
N-FHMM	0.000	0.000	0.000	0.000	0.000	0.000

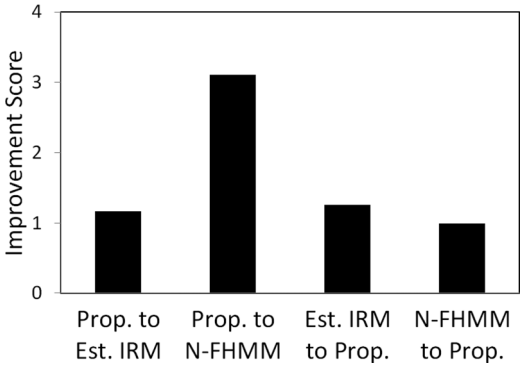

FIG. 6. Sound quality improvement scores when each signal is preferred. "Prop. to Est. IRM" indicates the quality improvement score when the subject indicates that the proposed signal has better quality than the signal produced with an estimated IRM.

The mean improvement scores for the different signals are shown in Fig. 6. Recall that the listener provides improvement scores for the signal that they believe is better in quality, where the score indicates the amount of quality improvement (see Sec. IV A). The average improvement score of the proposed system over the estimated IRM is approximately 1.2, indicating a slight but noticeable quality improvement. A quality improvement of roughly 3 is given for the proposed system over the N-FHMM, indicating that the quality is largely better. When the estimated IRM is preferred (only 25% rate), its quality improvement is approximately 1.3, indicating that it is slightly better than our proposed system. A quality improvement score of 1 is given for N-FHMM over our system, but this occurs at a rate less than 1%.

V. CONCLUDING REMARKS

We have proposed to use deep neural networks 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 DNN used for IRM estimation helps ensure that the NMF activations only approximate the speech component of the mixture, while the second DNN ensures that clean speech is estimated.

Our system is compared against similar two-stage approaches that combine masking with NMF reconstruction. The performance of these systems is limited since the NMF reconstruction stage uses speech models to approximate masked speech, while our system approximates clean speech. Traditional model-based approaches (i.e., supervised NMF and semi-supervised N-FHMM) are also compared to our method. The results presented in this paper show that our system improves perceptual quality at low

signal-to-noise ratios and it generally outperforms these comparison methods both objectively and in a listening evaluation.

ACKNOWLEDGMENTS

This research was supported in part by an AFOSR grant (No. FA9550-12-1-0130), an NIDCD grant (No. R01 DC012048), and the Ohio Supercomputer Center. A preliminary version of this paper, without the subject test, will be presented at ICASSP 2015 (Williamson *et al.*, 2015). We thank Gautham Mysore for assisting with our N-HMM and N-FHMM implementations.

Arehart, K. H., Kates, J. M., Anderson, M. C., and Harvey, L. O., Jr. (**2007**). "Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 1150–1164.

Duchi, J., Hazan, E., and Singer, Y. (**2010**). "Adaptive subgradient methods for online learning and stochastic optimization," J. Mach. Learn. Res. **12**, 2121–2159.

Eggert, J., and Korner, E. (**2004**). "Sparse coding and NMF," IEEE Conf. Neural Netw. **4**, 2529–2533.

Févotte, C., Bertin, N., and Durrieu, J-L. (**2009**). "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," Neural Comput. **21**, 793–830.

Han, K., Wang, Y., and Wang, D. L. (**2014**). "Learning spectral mapping for speech dereverberation," in *Proceedings of ICASSP*, pp. 4661–4665.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029–3038.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

ITU-R (**2001**). "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.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.

Koning, R., Madhu, N., and Wouters, J. (**2015**). "Ideal time-frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," IEEE Trans. Biomed. Eng. **62**, 331–341.

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

Mysore, G. J., and Smaragdis, P. (**2011**). "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proceedings of ICASSP*, pp. 17–20.

Mysore, G. J., Smaragdis, P., and Raj, B. (**2010**). "Non-negative hidden Markov modeling of audio with application to source separation," in *Proceedings of LVA/ICA*, pp. 1–8.

Nair, V., and Hinton, G. E. (**2010**). "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of International Conference on Machine Learning*, pp. 807–814.

Narayanan, A., and Wang, D. L. (**2013**). "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of ICASSP*, pp. 7092–7096.

Seung, H. S., and Lee, D. (**2001**). "Algorithms for non-negative matrix factorization," Adv. Neural Inf. Process. Syst. **13**, 556–562.

Smaragdis, P. (**2007**). "Convolutive speech bases and their application to supervised speech separation," IEEE Trans. Audio Speech Lang. Process. **15**, 1–12.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136.

Virtanen, T. (**2007**). "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," IEEE Trans. Audio Speech Lang. Process. **15**, 1066–1074.

Wang, Y., Han, K., and Wang, D. L. (**2013**). "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio Speech Lang. Process. **21**, 270–279.

Wang, Y., and Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio Speech Lang. Process. **21**, 1381–1390.

Wang, Y., Narayanan, A., and Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE Trans. Audio Speech Lang. Process. **22**, 1849–1858.

Williamson, D. S., Wang, Y., and Wang, D. L. (**2014a**). "A two-stage approach for improving the perceptual quality of separated speech," in *Proceedings of ICASSP*, pp. 7084–7088.

Williamson, D. S., Wang, Y., and Wang, D. L. (**2014b**). "Reconstruction techniques for improving the perceptual quality of binary masked speech," J. Acoust. Soc. Am. **136**, 892–902.

Williamson, D. S., Wang, Y., and Wang, D. L. (**2015**). "Deep neural networks for estimating speech model activations," in *Proceedings of ICASSP*, pp. 5113–5117.

Wilson, K., Raj, B., Smaragdis, P., and Divakaran, A. (**2008**). "Speech denoising using nonnegative matrix factorization with priors," in *Proceedings of ICASSP*, pp. 4029–4032.

Xu, Y., Du, J., Dai, L., and Lee, C. (**2014**). "An experimental study on speech enhancement based on deep neural networks," IEEE Sign. Process. Lett. **21**, 65–68.

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