A STRUCTURE-PRESERVING TRAINING TARGET FOR SUPERVISED SPEECH SEPARATION

Yuxuan Wang[1] *and DeLiang Wang*[1,2]

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

ABSTRACT

Supervised learning based speech separation has shown considerable success recently. In its simplest form, a discriminative model is trained as a time-frequency masking function, where the training target is an ideal mask. Ideal masks, such as the ideal binary masks, are structured spectro-temporal patterns. However, previous formulations do not model prominent output structure. In this paper, we propose an alternative training target that is explicitly related to mask structure. We first learn a compositional model of the square-root ideal ratio mask that is closely related to the Wiener filter. Instead of directly estimating the ideal mask values, we learn to predict the weights for resulting mask-level spectro-temporal bases, which are then used to generate the estimated masks. In other words, the discriminative model is used to predict the parameters of a generative model of the target of interest. Experimental results show consistent improvements in low SNR conditions by adopting the new training target.

Index Terms— Speech separation, deep neural networks, training target, spectro-temporal patterns

1. INTRODUCTION

Speech separation is a central problem in speech processing. Monaural speech separation segregates target speech from background noises using only one microphone. Despite many important applications, monaural speech separation remains a largely unsolved problem for decades. Monaural speech separation is particularly challenging when dealing with low signal-to-noise (SNR) ratio and non-stationary broadband noises. In these cases, speech information is typically buried in noise in the majority of time-frequency (T-F) units, rendering traditional methods, such as speech enhancement, ineffective.

Recently, speech separation has been formulated as a supervised learning problem (e.g., [7]) with success. In its simplest form, acoustic features are extracted from noisy mixtures after a time-frequency analysis. These features are used to train a discriminative model that attempts to predict some kind of ideal masks. In other words, the discriminative model acts as a masking function, where the inputs are noisy features and the outputs are mask values. The modeling power of machine learning techniques enables monaural separation in challenging conditions possible. In a recently conducted listening test [4], we have shown that by predicting the ideal binary mask (IBM) [15], a deep neural network (DNN) based monaural separation system significantly improves the intelligibility of noisy speech for hearing impaired listeners.

Previous supervised separation systems typically predict a mask value at each T-F unit independently. However, the ideal mask, whether binary or ratio, has strong spectro-temporal structure due to speech production mechanisms, which are largely ignored in previous systems. In fact, recent research has shown that it is the patterns in the IBM that carry important intelligibility information [9, 11]. Therefore, explicitly modeling the output structure in a learning algorithm will likely improve the performance in challenging conditions.

We propose the square-root ideal ratio mask (IRM) as the training target, which is different from previous systems that typically predict the IBM. Then, we propose to associate the training target of DNNs with the structure in the IRM. Specifically, we learn a compositional model to decompose the square-root IRM into a set of spectro-temporal bases and associated weights. Instead of directly estimating the IRM, the DNN is trained to estimate the weights that are used to linearly combine the mask bases to generate the estimated mask. By switching to this intermediate target, the training should be faster and the estimated mask is expected to be constructed from useful mask patterns.

Compositional models, in particular non-negative matrix factorization (NMF), have been used in source separation (e.g., [2, 13]). However, the proposed method is fundamentally different from NMF based methods. Supervised NMF typically models speech and noise spectra separately and finds a linear combination of them to fit the observed mixture spectra. In this sense, NMF can be loosely considered as a generative model. In contrast, our method is purely discriminative in the way that it learns a mapping from noisy features to the parameters (e.g., weights) of a compositional model, which serves as an intermediate step to obtaining the

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

estimated mask. The compositional model is applied to the ideal mask only to represent its spectro-temporal structure. In addition, it does not have to be an NMF.

2. SUPERVISED SPEECH SEPARATION

The computational goal of supervised speech separation is to estimate certain type of ideal masks capable of improving human speech intelligibility. We describe the training procedure used in this paper as follows. We use a 32-channel gammatone filterbank as our analysis frontend. The noisy mixtures are passed to the gammatone filterbank with center frequencies ranging from 50 to 8000 Hz. By windowing the filter response in each filter channel, a T-F representation called cochleagram [16] is formed for the noisy mixture. Acoustic features are then extraced and fed as input to a deep neural network, where the training target is provided by the ideal mask of interest.

The supervised learning framework offers us great flexibility in system design. When the ideal mask of interest is the IBM, the DNN is trained as a classifier to predict whether a T-F unit is target dominant or not. When the ideal mask is a ratio mask, the DNN is trained as a regressor to estimate the ideal gains. Using other training targets such as the instantaneous SNR is also possible [10]. In this paper, the baseline DNN is trained to predict the square root of the ideal ratio mask, which is defined as:

$$d(c,m) = \sqrt{\frac{S(c,m)^2}{S(c,m)^2 + N(c,m)^2}} \quad (1)$$

where $d(c,m)$ denotes the gain at channel c and time frame m. $S(c,m)^2$ and $N(c,m)^2$ denote the clean speech energy and noise energy at channel c and time frame m, respectively. As can be seen, this training target is closely related to the square-root Wiener filter, which is widely used in speech enhancement.

The baseline DNN system uses a window of features to predict the square-root IRM at each time frame. We use the combined acoustic features proposed in [17], including amplitude modulation spectrogram, relative spectral transform and perceptual linear prediction, mel-frequency cepstral coefficients, and gammatone filterbank power spectra.

3. STRUCTURE-PRESERVING TRAINING TARGET

Previous supervised speech separation systems directly learn a map from noisy features to the ideal mask of interest. This could be a difficult task especially at low SNR conditions, as features might be too noisy to be discriminative. Strong spectro-temporal correlations are fundamental to speech due to linguistic constraints and speech production mechanisms. This is well reflected in the ideal mask, and such output structure could be used to regularize the learning and to help the estimation.

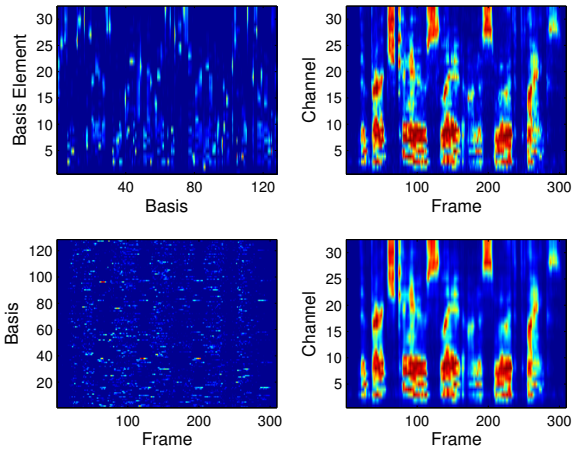

Fig. 1. A compositional model of the square-root IRM. Top left: 128 learned bases by training an NMF on square-root IRM (sliding window of 5 frames). Top right: the square-root IRM of a noisy mixture at -5 dB. Bottom left: basis weights inferred by the NMF. Bottom right: the reconstructed square-root IRM by using the resulting weights and bases.

Recent research has revealed the strong correlations between the patterns in the IBM and human intelligibility score as well as automatic speech recognition performance [9, 11]. Ideally, these patterns should have good correspondences to the underlying linguistic units such as phones or subphones. We assume that these patterns are constituents that can be combined to construct the ideal mask. In this paper, we learn an additive, compositional model of the ideal mask. Specifically, a simple NMF is trained on the square-root IRM and the resulting bases are considered as the constituent structures. To capture temporal structures, the NMF is trained on a window of frames instead of single time slices. After training, the resulting weights, bases, along with the NMF model form a generative model[1] of the square-root IRM. The weights are used to linearly combine the corresponding bases to reconstruct the original mask. Figure 1 shows the learned bases, weights and the reconstructed mask. We can see that the simple NMF is sufficient to reconstruct the square-root IRM.

Instead of directly estimating the square-root IRM, we propose to estimate the basis weights of its compositional model. As shown above, correctly estimated basis weights will lead to (almost) ideal masks. We expect that the estimation of basis weights tends to be more error tolerable than the direct estimation of masks per se. The rationale is that even when some weights for a particular frame are erroneous,

[1] Strictly speaking, the standard NMF is not a generative model as it is not probabilistic. Here the term is slightly abused to refer to models that can produce/reconstruct the data of interest.

Table 1. STOI measure comparisons between different systems and training targets

System	Target	Babble			Factory			Speech-shaped		
		-5 dB	-2 dB	0 dB	-5 dB	-2 dB	0 dB	-5 dB	-2 dB	0 dB
Mixture	n/a	0.562	0.629	0.676	0.571	0.635	0.681	0.607	0.675	0.722
DNN	Square-root IRM	0.644	0.745	0.789	0.703	0.776	0.817	0.770	0.828	0.859
DNN	Mask basis weight	**0.660**	**0.755**	**0.800**	**0.721**	**0.786**	**0.823**	**0.775**	**0.831**	**0.862**
Supervised NMF	Spectrogram	0.596	0.667	0.716	0.632	0.696	0.738	0.682	0.752	0.789
Hendriks et al. [5]	DFT magnitudes	0.497	0.586	0.645	0.529	0.613	0.669	0.604	0.681	0.731
DNN	(Spectrogram) basis weight	0.678	0.753	0.790	0.730	0.787	0.816	0.783	0.826	0.850

Table 2. PESQ score comparisons between different systems and training targets

System	Target	Babble			Factory			Speech-shaped		
		-5 dB	-2 dB	0 dB	-5 dB	-2 dB	0 dB	-5 dB	-2 dB	0 dB
Mixture	n/a	1.396	1.573	1.711	1.254	1.428	1.540	1.497	1.648	1.735
DNN	Square-root IRM	1.625	1.862	2.105	1.761	2.075	2.246	1.934	2.168	2.350
DNN	Mask basis weight	1.668	1.953	2.246	1.860	2.136	2.273	1.970	2.193	2.360
Supervised NMF	Spectrogram	1.462	1.631	1.752	1.511	1.688	1.832	1.653	1.836	1.961
Hendriks et al. [5]	DFT magnitudes	1.238	1.490	1.668	1.368	1.653	1.852	1.475	1.732	1.921
DNN	(Spectrogram) basis weight	1.670	1.913	2.070	1.870	2.073	2.200	1.981	2.163	2.255

the rest of the correctly estimated weights can still contribute useful structure, which may comprise to make part of the estimated mask correct. In other words, training to estimate the basis weights helps preserve the structure in the final algorithm output, which may lead to more perceptually relevant results.

4. EXPERIMENTS

4.1. Experimental Settings

We use the IEEE sentences recorded by a male speaker as the speech corpus [6]. Three challenging broadband noises, i.e., a speech-shaped noise (SSN), a factory noise and a babble noise are additively mixed with clean speech to create the training and test mixtures. To create the training set, we mix 200 sentences with the first two minutes of each of the three noises at -5, -2 and 0 dB. Each clean sentence is mixed with 20 randomly picked noise segments, creating a training set of 4000 mixtures for each type of noise. To create the test set, we use 50 unseen sentences to mix with random cuts from the last two minutes (no overlap with any training noise segments) of the three noises, also at -5, -2 and 0 dB. The SSN is stationary, whereas the factory and babble noise are highly non-stationary. We point out that these noises are extremely difficult to separate at negative SNR conditions. For example, the human (normal-hearing) intelligibility score for the used factory noise at -5 dB is well below 50% [8].

We use the standard feed-forward DNNs (multi-layer perceptrons) as the discriminative model throughout all the experiments. All DNNs use three hidden layers, each having 1024 rectified linear units. The networks are discriminatively trained using the standard backpropagation algorithm with dropout regularizations, and no unsupervised pretraining is used. Adaptive stochastic gradient descent (AdaGrad) is used as the optimizer [1]. We use a window (5 frames) of combined features as inputs to the DNN (input dimension is 1230). To compare the standard and proposed training targets, we train two sets of DNNs. The first set uses a 32-D output layer to directly estimate the square-root IRMs across all 32 channels. The standard NMF is used to learn 128 bases of the square-root IRM using a sliding window of 5 frames. Therefore, the second set uses a 128-D output layer to estimate the weights of these bases. We use the standard sigmoid activation functions for the output layers of the mask-estimating DNNs, as the square-root IRM is bounded within $[0, 1]$. The values of the basis weights are greater than or equal to 0 (due to non-negativity) but are not necessarily less than or equal to 1. To accommodate this, we use bounded linear output units for weight-estimating DNNs, which are defined as $f(x) = \max(0, \min(x, m))$. Here m denotes the largest weight value found in the training set. Note that we also used the bounded linear units (with $m = 1$) for mask-estimating DNNs but did not achieve better results. This indicates that the choice of output activation functions does not contribute to the performance differences shown next.

4.2. Results

To put the performance of supervised speech separation in perspective, we also compare with NMF and speech enhancement based systems. We compare with supervised NMF [18], where the speech bases and noise bases are trained separately for each type of noise using exactly the same training data used by DNNs. We have made efforts to tune its perfor-

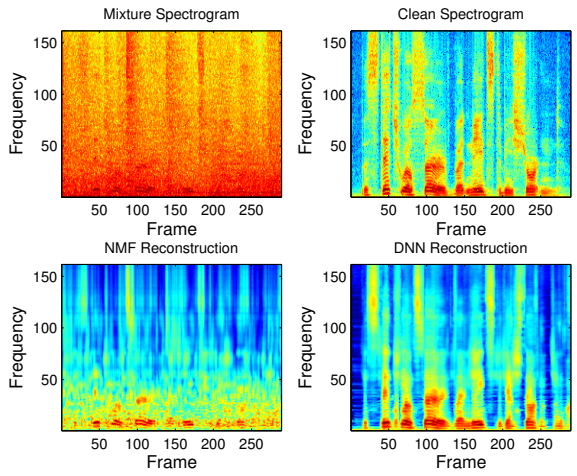

Fig. 2. Spectrogram reconstruction of a -5 dB mixture (factory noise). The DNN is trained to estimate the weights of the spectrogram bases learned by an NMF.

mance, and the best NMF results are obtained by using a sliding window of 11 frames [2]. We use 80 and 160 bases to represent speech and noise spectrogram, respectively. We also tried using convolutive NMF [13] to learn more invariant spectro-temporal bases, but did not achieve betters results for the noises used in this paper. For speech enhancement, we compare with the algorithm proposed by Hendriks et al. [5], which is considered as a state-of-the-art in the speech enhancement community. In evaluation, we use Short-Time Objective Intelligibility measure [14] (STOI) and Perceptual Evaluation of Speech Quality score [12] (PESQ) to evaluate the objective speech intelligibility and speech quality improvement, respectively. Both STOI and PESQ are obtained by comparing with clean speech.

STOI and PESQ results for different systems with different training targets are presented in Table 1 and 2. Across all SNR conditions, switching training target from mask to mask basis weight gives consistent improvements. Note that STOI represents a correlation between 0 and 1, and a 1% absolute improvement is considered significant. For example, using the conversion formula for the IEEE corpus given in Table II of [14], the projected intelligibility improvement is about 4% in the case of -5 dB babble. Estimating weights seems to be more helpful in lower SNR and non-stationary noise cases. The STOI improvement is only marginal in the case of SSN. Switching targets gives slight but also consistent improvements in PESQ. Although using exactly the same training data, supervised NMF is substantially worse than DNNs in both STOI and PESQ (perceptually it is also much worse). In low SNR conditions, data-driven techniques seem to be very important, which is reflected by the comparisons with Hendriks et al.'s system. In fact, except SSN, the STOI results of Hendriks et al.'s system are significantly worse than those of unprocessed. Although designed to improve speech quality, it is also worth noting that NMF and speech enhancement are significantly worse than masking based DNN approaches in terms of PESQ.

4.3. Discussions

The proposed method can be easily extended. The ideal target of interest can vary according to applications, and is certainly not limited to mask estimation. For example, we can learn spectrogram bases and use DNNs to predict corresponding weights, which along with the pre-learned bases can be used to directly reconstruct clean spectrograms. When the bases are learned by NMF, this can be thought as a supervised and much more non-linear fashion to perform NMF inference. An example of reconstructing the clean spectrogram of a -5 dB mixture (factory noise) is shown in Figure 2, where we can see that the proposed method is much better than a supervised NMF that uses the same training data (thus the clean spectrogram bases are the same for both DNN and NMF). In a few cases, spectrogram reconstruction offers further improvements, which can be seen from the last row in Table 1 and 2. This seems to be an interesting future work for us. Enabled by the supervised learning framework, the proposed method can also be easily applied to other applications, such as bandwidth extension and dereverberation.

5. CONCLUSIONS

Choosing a suitable training target is important for supervised learning. It is possible that predicting an intermediate target makes learning easier and generalize better [3]. Motivated by the spectro-temporal pattens in the ideal masks, we have proposed a structure-preserving training target as an alternative to directly estimating the mask values. Switching to this intermediate target provides consistent STOI and PESQ improvements, especially for non-stationary noises in very low SNR conditions. In addition, we have also demonstrated that a standard DNN that estimates the square-root IRM can perform substantially better than supervised NMF and speech enhancement in low SNR conditions.

The supervised learning framework is very flexible that it enables the proposed method to be applicable to many kinds of target of interest and/or to many applications other than speech separation. We want to point out that the proposed method does not rely on NMF. Any kind of generative models, or even speech production models, should be applicable. These are all interesting future work.

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

6. REFERENCES

[1] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, pp. 2121–2159, 2011.

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

[3] C. Gulcehre and Y. Bengio, "Knowledge matters: Importance of prior information for optimization," in *International Conference on Learning Representations (ICLR)*, 2013.

[4] E. Healy, S. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *Journal of the Acoustical Society of America*, pp. 3029–3038, 2013.

[5] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. ICASSP*, 2010, pp. 4266–4269.

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

[7] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 625–638, 2009.

[8] 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*, pp. 1486–1494, 2009.

[9] U. Kjems, J. Boldt, M. Pedersen, T. Lunner, and D. Wang, "Speech intelligibility in background noise with ideal binary time-frequency masking," *Journal of the Acoustical Society of America*, vol. 126, pp. 1415–1426, 2009.

[10] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7092–7096.

[11] ——, "The role of binary mask patterns in automatic speech recognition in background noise," *Journal of the Acoustical Society of America*, pp. 3083–3093, 2013.

[12] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

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

[14] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 2125–2136, 2011.

[15] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Divenyi P., Ed. Kluwer Academic, Norwell MA., 2005, pp. 181–197.

[16] D. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.

[17] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 270–279, 2013.

[18] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. ICASSP*, 2008, pp. 4029–4032.