ROBUST SPEECH RECOGNITION FROM RATIO MASKS

Zhong-Qiu 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
{wangzhon, dwang}@cse.ohio-state.edu

ABSTRACT

Robustness against noise is crucial for automatic speech recognition systems in real-world environments. In this paper, we propose a novel approach that performs robust ASR by directly recognizing ratio masks. In the proposed approach, a deep neural network (DNN) is first trained to estimate the ideal ratio mask (IRM) from a noisy utterance and then a convolutional neural network (CNN) is employed to recognize estimated IRMs. The proposed approach has been evaluated on the TIDigits corpus, and the results demonstrate that direct recognition of ratio masks outperforms direct recognition of binary masks and traditional MMSE-HMM based method for robust ASR.

Index Terms— Robust ASR, Ideal Ratio Mask, Ideal Binary Mask, CNN, DNN

1. INTRODUCTION

The performance of traditional speech recognition systems degrades substantially in noisy environments, which is largely due to the mismatch between training and test conditions such as different background noises, different channels, different speaker characteristics and different input SNRs. Many methods are proposed in the past few years [8]. Feature domain methods extract robust speech features such as RASTA-PLP [2] or de-noise the noisy speech first before recognition. Model domain methods [5] change the parameters of a speech recognizer to account for the effect of distortions. Other methods such as SPLICE [1] utilize prior knowledge about the distortions based on so-called stereo data and reconstruct the clean speech from noisy speech. Methods such as VTS [9] explicitly model the effect of noise and channel distortions on clean speech for model adaptation and distortion estimation. Although many methods are proposed, a lot of research remains to be done to make ASR robust in real-world environments.

Recently, supervised speech separation (e.g. [20]) has shown considerable potential as a front end for robust speech recognition [19][10][18]. These methods typically estimate the ideal binary mask (IBM) – a binary T-F mask that identifies speech dominant and noise dominant T-F units, or the ideal ratio mask (IRM) – a ratio T-F mask that represents the ratio of speech energy to the mixture energy within each T-F unit. An estimated mask is then used as a front-end to enhance the noisy speech [11][7]. Afterwards, the enhanced speech is recognized by a recognizer trained using enhanced or noisy speech. One motivation for using a masking based method is that the IBM itself seems to encode adequate phonetic information for speech recognition [16]. Based on this insight, Narayanan and Wang [12] proposed a new recognition method by directly recognizing an estimated IBM of a noisy utterance. This highly different method gives significant improvements over an MMSE-HMM based method, especially at low input SNR conditions.

Clearly a ratio mask contains more information than a binary mask. Does the IRM contain more phonetic information for robust speech recognition than the IBM? Are there performance differences in estimated IBM and estimated IRM for ASR?

In this study, we extend the Narayanan and Wang technique [12] to the ratio masking domain, and recognize ratio masks as visual patterns for the purpose for ASR. In addition, we employ a DNN for mask estimation rather than a traditional CASA method. Our method yields significant improvements in recognition rate on the TIDigits corpus [6]. The superior performance of ratio mask recognition over binary mask recognition suggests that ratio masking may be more suitable for robust ASR than binary masking.

2. SYSTEM DESCRIPTION

The key idea behind our system is the direct recognition of estimated IRMs of noisy utterances as visual patterns. Figure 1 shows the typical IRMs for 11 noisy isolated digit utterances (0-9 and ‘oh’) along with the corresponding IBMs. We can see that both IRMs and IBMs have distinct visual patterns. In addition, the IRM contains more information than the IBM because of its continuous values. The proposed method uses DNN to estimate the IRM and then uses CNN to recognize estimated IRMs.

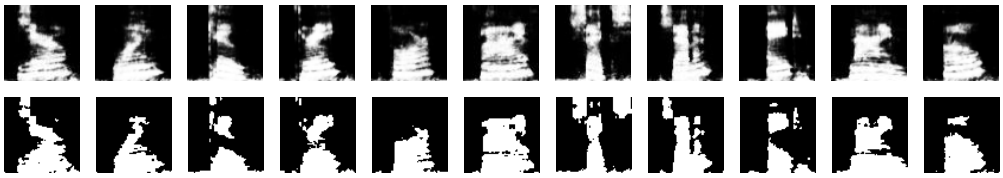


Figure 1. Comparison between typical IRMs (upper row) and IBMs (lower row) of 11 noisy digit utterances 0-9 and 'oh', arranged from left to right. These noisy utterances are created by mixing each clean single digit utterance with 32-speaker babble noise at 6 dB. The LC for calculating IBMs is set to 0 dB.

2.1 IBM and IRM

The IBM is a T-F mask calculated from premixed clean speech and noise. The clean speech and noise, scaled to a specific input SNR, are passed through a 64-channel gammatone filterbank with central frequencies ranging from 50 Hz to 8000 Hz on the equivalent rectangular bandwidth rate scale. The resulting signal of each channel is then divided into 20-ms frames with 10-ms overlap, producing a cochleagram for clean speech and noise, respectively [15]. For each T-F unit in the IBM, its value is set to 1 if the instantaneous SNR within that T-F unit is greater than a pre-defined local SNR criterion (LC) and 0 otherwise. Quantitatively, the IBM is defined [14] as

$$\mathrm{IBM}(t,f) = \begin{cases} 1, if\ SNR(t,f) > LC \\ 0, otherwise \end{cases}$$

The LC is set to 0 dB in our experiments.

The definition of the IRM is similar [13]. Rather than binary values, the IRM uses the ratio of speech energy over mixture energy within each T-F unit, i.e.

$$\mathrm{IRM}(t,f) = \frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)} = \frac{10^{SNR(t,f)/10}}{10^{SNR(t,f)/10} + 1}$$

where $S^2(t,f)$ and $N^2(t,f)$ denote the speech energy and noise energy at a particular T-F unit, respectively.

2.2 Mask Estimation

At the test stage, we only have mixtures of clean speech and noises, so we need to estimate the mask of each mixture. We use a DNN for mask estimation, which has been successfully applied to supervised speech separation in the past few years [20]. The diagram of mask estimation in our approach is shown in Figure 2. In our experiments, all DNNs have two hidden layers and each hidden layer has 1024 sigmoid units. In the output layer, there are 64 sigmoid units which correspond to the number of frequency channels. No pre-training is used in our experiments. The dropout ratio of the hidden layers and input layer is set to 0.3 and 0.1, respectively. The maximum L2 norm of the incoming weights of each neuron is chosen to be 10. Standard back-propagation algorithm is used to train the network with mini-batch size 1024. The maximum number of epochs is set to 200. The learning rate is linearly decreased from 1 to 0.001. The momentum is increased from 0.5 to 0.95 in the first 60 epochs and kept fixed at 0.95 afterward. The network is trained to minimize the mean square error frame-wisely, and the labels for training DNN come from ideal masks. Note that these parameters are selected to minimize the validation error of IBM estimation on a validation set. When training DNN for IRM estimation, we use exactly the same parameters to facilitate comparison.

The features we use for mask estimation consist of a complementary frame-level feature set [17] and its delta components. The complementary feature set contains 31 dimensional mel-frequency cepstral coefficients (MFCC), 64 dimensional gammatone filterbank power spectra (GF), 13 dimensional relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP) and 15 dimensional amplitude modulation spectrogram (AMS). To further incorporate temporal context, we splice a five-frame window as the input to the DNN. So in our approach, the input feature dimension is 1230 ((31+64+13+15)x2x5), and the output dimension is 64 which corresponds to the number of filter channels in one frame.

After obtaining an estimated mask of a noisy utterance, its centroid is calculated, and then used to extract a 64x64 image for later recognition by choosing 32 frames on the left and 31 frames on the right side of the centroid. Note that a 64-frame window is longer than all the single digit utterances in the TIDigits corpus.

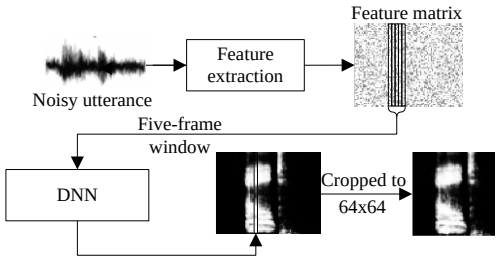


Figure 2. Diagram of DNN based mask estimation

2.3 Mask Recognition

CNN has shown considerable success in digit recognition from gray-scale images. Since CNN can capture local topology and is relatively invariant to small shift and distortion in an image, it is very suitable for image

recognition [3]. In this study, we employ CNN to recognize estimated masks for three reasons. First, the ratio masks as shown in Figure 1 clearly show characteristic visual patterns. Second, generally speaking, the centroid of an estimated mask cannot be calculated perfectly, so translational invariance is a desired property to overcome this issue. Third, there are substantial distortions in the masks for different types of noises and input SNR conditions. For these reasons, CNN is chosen for our mask recognition.

The architecture of CNN in our experiments is shown in Figure 3. It follows the CNN used in [12] for binary mask recognition and is similar to LeNet5 [4]. Layer C1, C3 and C5 are convolutional layers with kernel size 5x5, 6x6 and 5x5, respectively. S2 and S4 are mean pooling layers with kernel size 3x3. The output layer is fully connected to the preceding layer and has 11 units which correspond to the 11 ratio patterns we want to recognize. The feature map size of C1, C3 and C5 are set to 7, 20 and 150, respectively. The network is trained using the stochastic diagonal Levenberg-Marquardt algorithm for 20 epochs. A validation set is used for early stopping and parameter tuning.

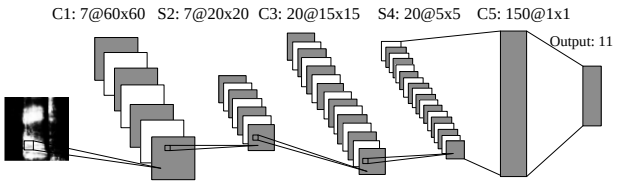


Figure 3. The architecture of CNN for mask recognition

In addition, as in LeNet5, the connections between S2 and C3 are set to incomplete connections, which force different feature maps to extract different and complementary features [4]. The connections between S2 and C3 in our experiments are shown in Table 1, where 'X' denotes a connection between these two feature maps. In our experiments, we find that using incomplete connections between S2 and C3 brings us 7% improvement in recognition rate than using complete connections.

Table 1. Incomplete connections between layer S2 and C3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	X				X	X	X	X		X	X			X	X	X			X	X
2	X	X				X	X	X			X	X				X	X		X	X
3	X	X	X				X		X	X		X	X		X			X	X	X
4	X	X	X	X					X	X	X		X	X		X	X	X		X
5		X	X	X	X			X			X	X		X	X			X	X	X
6			X	X	X	X		X	X			X	X			X	X			X
7				X	X	X	X		X	X			X	X	X		X	X		X

3. EXPERIMENTAL SETTINGS

In order to make a direct comparison with the binary mask recognition approach [12], we use a similar experimental setup. A single digit utterance subset of the TIDigits corpus is utilized to evaluate our proposed method. Our training and test utterances consist of all the single digit utterances of 55 male speakers and all the single digit utterances of 56 different male speakers, respectively. There are 11 digits (0-9 and 'oh') in total. Each digit is spoken by each speaker for two times. We use speech shape noise, 32-speaker babble noise and cocktail party noise as our training and test noises. The latter two noises are non-stationary.

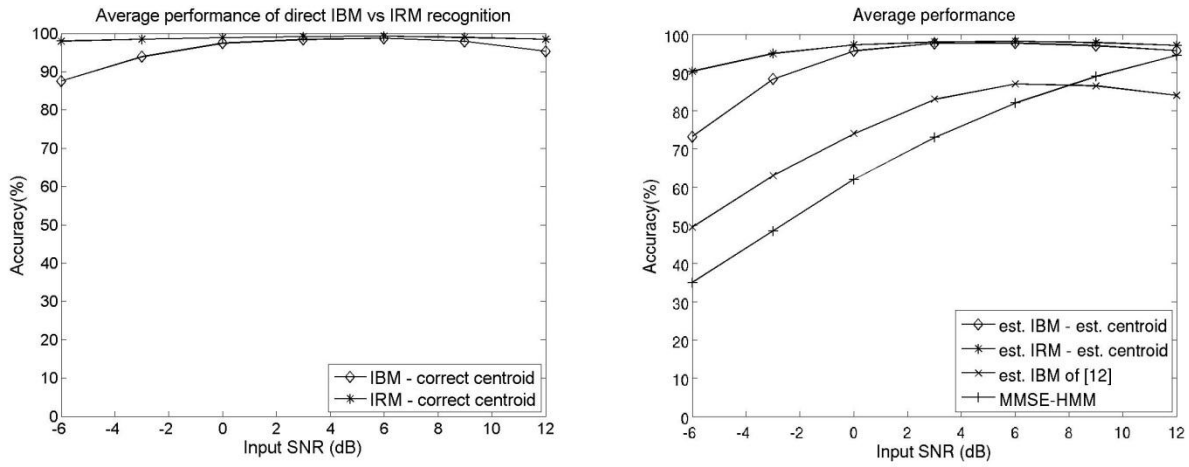
The DNN based ratio mask estimator in our system is trained through multi-condition training. The training set for mask estimation is constructed by mixing each clean training utterance with each noise at -6, -3, 0, 3, 6, 9 and 12 dB. A validation set which contains the noisy utterances of five randomly selected speakers is kept separate from the training set for parameter tuning and early stopping.

The CNN based mask recognizer in our system is trained using IRMs created by mixing each clean training utterance with each noise only at 6 dB. Before training, each IRM is cropped into a 64x64 image by using the center of speech range as the centroid. When testing, the centroid of each estimated mask is calculated using the method described earlier. A validation set which contains the IRMs from the same five speakers as used in the validation set of mask estimation is kept separate from the training set for parameter tuning and early stopping.

The test set for our overall system is constructed by mixing each clean test utterance with each noise at -6, -3, 0, 3, 6, 9 and 12 dB. By including different input SNR conditions in the test set, we can test the generalization ability of the CNN based mask recognizer.

4. RESULTS AND DISCUSSIONS

We first report the results of directly recognizing cropped IRMs using correct centroids (center of the speech range) based on the CNN trained on IRMs, and this will give us the ceiling performance of our proposed method. Figure 4(a) shows that the average recognition rates of all the noise types at all the seven input SNR conditions are greater than 98%. In the same figure, we also present the result of directly recognizing cropped IBMs with correct centroids on the same noisy utterances using the same CNN but trained on IBMs. Our obtained result on IBM recognition is very close to the result reported in Fig. 3(d) of [12]. By comparison, we can see that the IRM based result is better than the IBM based result. For these two approaches, the best recognition rate occurs when input SNR is at 6 dB, which is expected since it is the same as the training SNR of the CNN based mask recognizer. When input SNR increases or decreases, the performance for both recognizers drops. This drop is largely due to the mismatch between training



(a) (b)

Figure 4. Average recognition results of different noises at different input SNRs. (a). Average results of directly recognizing IRMs and IBMs with correct centroids using CNN. (b). Average results of recognizing estimated IRMs and estimated IBMs with estimated centroid using CNN, the method used in [12] and the MMSE-HMM approach.

and testing SNR. Take the IBM as example. When input SNR increases from 6 dB, more and more 1's will occur in the IBM. Similarly, when input SNR decreases from 6 dB, there will be fewer and fewer 1's. As a result, test IBMs will be different from training IBMs. We can also see from Figure 4(a) that, during testing, the IBM recognizer suffers more from the SNR mismatch. This is probably because, for the IRM, the value of each T-F unit is between 0 and 1 rather than binary, therefore more discriminative structure would be retained than by the IBM recognizer, especially in low input SNR conditions.

In Figure 4(b), we report the result of recognizing cropped estimated IRMs with estimated centroids. Note that this is a realizable system. As shown in Figure 4(b), the recognition rate is above 95% when input SNR is greater than or equal to -3 dB. The performance is still greater than 90% when input SNR drops to -6 dB.

We now compare our approach with the binary mask based approach, both using estimated centroids when cropping. We can see that our method outperforms the one in [12] in all input SNR conditions. As in the ideal mask case, the performance gap is greater when input SNR decreases. Besides the reason mentioned before, the larger gap may reflect the fact that, for IBM estimation, a threshold (0.5 in this study) needs to be set to make final binary prediction, which may lead to loss of discriminative information for IBM estimation. For IRM estimation, there is no need to set a hard threshold. Figure 4(b) also shows the corresponding results by Narayanan and Wang and a traditional MMSE-HMM approach on the same dataset (Figure 3(d) in [12]). In the MMSE-HMM system, there are 12 word level models (0-9, oh and silence), each with 8 states. Every state emission probability is modeled as a mixture of 10 Gaussians. The features for training are the MFCC feature extracted from clean utterances. The language model is designed to only allow one digit in an utterance. At the test stage, noisy utterances are first enhanced using the MMSE algorithm before decoding. The recognition rates of MMSE-HMM are close to but lower than those of the binary mask based method when input SNR is 12 dB. At other lower input SNR conditions, its performance is significantly worse than both of our ratio and binary mask methods. As shown in Figure 4(b), the DNN based IBM estimator in this study also brings significant improvements over the traditional CASA estimator used in [12].

5. CONCLUDING REMARKS

In this study, we have proposed a novel method to perform robust ASR by directly recognizing estimated IRMs of noisy utterances as visual patterns. The experiments on the TIDigits corpus suggest that ratio masks encode more useful information for robust ASR than binary masks, especially in low input SNR conditions.

Spoken digit recognition is not a challenging task in ASR, and future work needs to address more challenging robust recognition problems. Nonetheless, the findings in our study indicate that, at a minimum, ratio masks obtained from supervised speech separation likely contain discriminative information not exploited in traditional methods of robust ASR, and hence can complement these methods for further performance improvements.

6. ACKNOWLEDGEMENTS

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NSF grant (IIS-1409431), and the Ohio Supercomputer Center.

7. REFERENCES

[1] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments.," in *Proceedings of Interspeech*, 2000, pp. 806–809.

[2] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] Y. LeCun and L. Bottou, "Gradient-based learning applied to document recognition," *Proceedings of IEEE*, vol. 86, pp. 2278–2324, 1998.

[5] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, pp. 171–185, 1995.

[6] R. Leonard, "A database for speaker-independent digit recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1984, vol. 9, pp. 328–331.

[7] B. Li and K.C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1296–1305, 2014.

[8] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 745–777, 2014.

[9] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996, pp. 733–736.

[10] A. Narayanan, A. Misra, and K. Chin, "Large-scale, sequence-discriminative, joint adaptive training for masking-based robust asr," in *Proceedings of Interspeech*, 2015, pp. 3571–3575.

[11] A. Narayanan and D.L. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 826–835, Apr. 2014.

[12] A. Narayanan and D.L. Wang, "Robust speech recognition from binary masks," *The Journal of the Acoustical Society of America*, vol. 128, pp. 217–222, 2010.

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

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

[15] D.L. Wang and G.J. Brown, *Computational auditory scene analysis: principles, algorithms, and applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.

[16] D.L. Wang and U. Kjems, "Speech perception of noise with binary gains," *The Journal of the Acoustical Society of America*, vol. 124, pp. 2303–2307, 2008.

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

[18] Y. Wang, A. Misra, and K. Chin, "Time-frequency masking for large scale robust speech recognition," in *Proceedings of Interspeech*, 2015, pp. 2469–2473.

[19] Z.-Q. Wang and D.L. Wang, "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition," in *Proceedings of Interspeech*, 2015, pp. 2839–2843.

[20] Y. Wang and D.L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381–1390, 2013.