IDEAL RATIO MASK ESTIMATION USING DEEP NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION

Arun Narayanan* *and DeLiang Wang**†

*Department of Computer Science and Engineering
†Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
{narayaar,dwang}@cse.ohio-state.edu

ABSTRACT

We propose a feature enhancement algorithm to improve robust automatic speech recognition (ASR). The algorithm estimates a smoothed ideal ratio mask (IRM) in the Mel frequency domain using deep neural networks and a set of time-frequency unit level features that has previously been used to estimate the ideal binary mask. The estimated IRM is used to filter out noise from a noisy Mel spectrogram before performing cepstral feature extraction for ASR. On the noisy subset of the Aurora-4 robust ASR corpus, the proposed enhancement obtains a relative improvement of over 38% in terms of word error rates using ASR models trained in clean conditions, and an improvement of over 14% when the models are trained using the multi-condition training data. In terms of instantaneous SNR estimation performance, the proposed system obtains a mean absolute error of less than 4 dB in most frequency channels.

Index Terms— Computational Auditory Scene Analysis, instantaneous SNR, noise robust ASR, Aurora-4

1. INTRODUCTION

Noise robust speech recognition is a widely studied research problem with important practical applications [1]. Several methods aim to extract robust features like RASTA PLP [2] and AFE [3]; but merely tuning feature extraction has achieved limited success. Therefore, techniques like model adaptation and feature enhancement are commonly used. Adaptation techniques, like MLLR [4] and Vector Taylor series (VTS) based adaptation [5, 6], try to modify the model parameters to match the test conditions better. Such methods are computationally expensive, and may additionally need adaptation data. In contrast, feature enhancement techniques try to remove noise from a given mixture without modifying the model parameters. Such methods are, therefore, computationally more efficient. Examples of feature enhancement techniques include missing feature reconstruction [7], Wiener filtering [8], and VTS-based enhancement [9].

A popular way to perform feature enhancement is by using computational auditory scene analysis (CASA) based algorithms to perform speech separation prior to recognition. Inspired by the remarkable robustness of human listeners, CASA aims to develop speech separation algorithms motivated by the principles of auditory scene analysis [10]. A main goal of CASA is to estimate the ideal binary mask (IBM) [11], which identifies each unit in a time-frequency (T-F) representation of the noisy signal as speech dominant or noise dominant. With the IBM as the computational goal, the task of separation reduces to a binary classification problem. The IBM has been used for performing feature enhancement (or noise suppression) in ASR by either using the direct masking approach [12] or by performing reconstruction [7]. In direct masking, the IBM is used as a binary gain function to attenuate the energy within the noise-dominant T-F units. In reconstruction, the speech energy within the noise dominant units is estimated using the information available in the speech dominant units.

The performance of both the above methods depend largely on the quality of IBM estimation. Supervised classification-based algorithms have been used to perform the task of IBM estimation for speech separation [13, 14]. Such algorithms extract features at the T-F unit level, and perform classification using learning machines like SVMs and deep neural networks (DNN). One of the goals of this study is to evaluate performance of such algorithms on a robust ASR task. In robust ASR, it has been noted in earlier studies that estimating the ideal *ratio* mask may result in better performance [15][1]. Therefore, we also study: 1) how can such supervised learning algorithms be adapted to estimate the IRM and 2) the potential of such algorithms in improving noise robust ASR performance.

The rest of the paper is organized as follows. In Section 2, we discuss prior work related to IRM estimation. Section 3 provides the system description. Evaluation results are presented in Section 4. We conclude in Section 5.

2. PRIOR WORK

Soft masks have been used in several robust ASR studies [7, 16]. The values in a soft mask represent the probability of a T-F unit being speech dominant, and are typically used in a missing data framework to perform recognition. The masks are estimated either by applying a sigmoid function to the estimated *a priori* signal-to-noise-ratio (SNR) [16], or by using a Gaussian mixture model of speech to directly predict the posterior probability [7]. In an alternative approach, Srinivasan *et al.* estimate the IRM by learning the relationship between the binaural cues of interaural time and level differences, and the instantaneous SNR [15]. Note that instantaneous SNR is directly related to the IRM value at each T-F unit. They use the estimated IRM to perform feature enhancement and report improvements in ASR performance over using the estimated IBM.

The research described in this paper was supported in part by an AFOSR grant (FA9550-12-1-0130).

[1] The IBM can be thought of as a binary approximation to the IRM.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

Recently, van Hout and Alwan propose to estimate a smoothed ratio mask using noise power estimators and a median filter, which they use to perform feature enhancement in the log Mel spectral domain before cepstral transformation [17].

SNR estimation, which is a general task, has been widely studied in the context of speech enhancement. Typical algorithms estimate the *a priori* SNR which is used to obtain the gain at each T-F unit for enhancement [8]. A supervised learning algorithm to estimate the instantaneous SNR was proposed by Tchorz and Kollmeier [18]. Their system uses amplitude modulation spectrograms (AMS) as features and multi-layer perceptron (MLP) as the function estimator.

3. SYSTEM DESCRIPTION

The proposed system uses a supervised learning algorithm to estimate the IRM. The following subsections describe how the desired target is set, what features are used, and how the mapping function is learned.

3.1. Target signal

Mathematically, the ideal ratio mask, which is closely related to the Wiener gain, is defined as follows:

$$\begin{aligned} IRM(m,c) &= \frac{10^{(SNR(m,c)/10)}}{10^{(SNR(m,c)/10)}+1}, \\ \text{and } SNR(m,c) &= 10\log_{10}(x(m,c)/n(m,c)). \end{aligned}$$

Here, $x(m,c)$ and $n(m,c)$ denote the instantaneous speech and noise energy, respectively, at time frame m and frequency channel c. $SNR(m,c)$ denotes the instantaneous SNR in dB. Instead of directly estimating the IRM, our system estimates the instantaneous SNR transformed using a tunable sigmoid function:

$$d(m,c) = \frac{1}{1+\exp(-\alpha(SNR(m,c)-\beta))}. \quad (1)$$

$d(m,c)$ denotes the desired target while training. α controls the slope of the sigmoid, and β is the bias. By tuning α and β, we can control the range of SNR to focus on while training the system. In our experiments we set α to roughly have a 35 dB SNR span[2] centered at β, which is set to -6 dB. β corresponds to the threshold commonly used to define the IBMs [19]. The SNR to target mapping based on these chosen values is shown in Fig. 1.

During testing, the output of the system is mapped back to the corresponding IRM values so that they can be used as a filter to perform noise suppression.

3.2. Features

We perform mask estimation in the Mel spectral domain, which is a commonly used front-end to perform feature enhancement for ASR. To extract features, the pre-emphasized input signal is first filtered using a 26-channel Mel filterbank that spans frequencies from 50 Hz to 7 kHz. The filterbank is implemented using sixth order butterworth filters. The filter output in each channel is then used to extract the following T-F unit level features: 13 dimensional RASTA filtered PLP cepstral coefficients with delta and acceleration components, 31 dimensional Mel frequency cepstral coefficients (MFCC), 15 dimensional AMS features, and 6 dimensional pitch-based features along with their time and frequency delta components. While calculating these features the hop size is set to 10 msec; the frame size depends on the feature type – 20 msec frames are used for RASTA PLPs, MFCCs, and pitch-based features, and 32 msec frames are used for AMS features (see [20] for detailed descriptions of how these features are extracted). We use this group of features as it has been found to work well for IBM estimation [20].

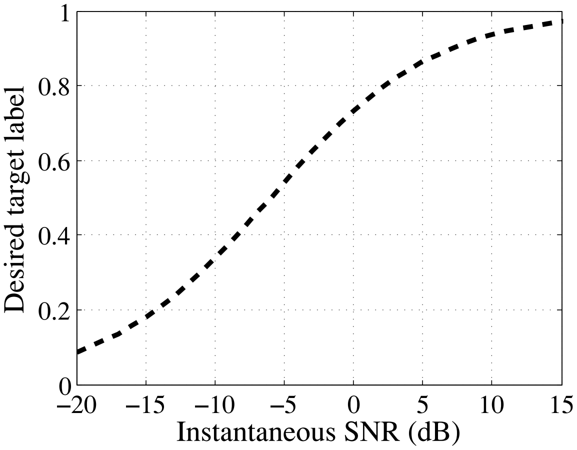

Fig. 1. The instantaneous SNR mapping function that is used to set the desired target during training.

3.3. Supervised learning

Following the supervised IBM estimation algorithm proposed in [21], we use deep neural networks to learn the function that maps the extracted features to the desired target (see Eq. 1). We take a two stage approach. In the first stage, 26 DNNs are trained, one for each frequency channel, using the features described above. The DNN training schedule includes an unsupervised pre-training phase and a supervised back-propagation phase, each consisting of 100 epochs [22]. The cross-entropy learning criterion is used during back-propagation. Each DNN has 103 input nodes corresponding to the feature dimensionality, 2 hidden layers each with 200 nodes, and an output layer with 1 node.

The DNNs learn the function using locally obtained features at each T-F unit, and do not directly make use of the information available in the neighboring units. Therefore, in the second stage, we learn MLPs with 1 hidden layer to smooth the output of the DNN. The MLPs are also trained for each frequency channel. They use the output of the DNNs in a neighborhood surrounding each T-F unit as input, and are trained to re-estimate the same targets as the DNNs. The neighborhood, which was chosen based on ASR performance on a development set, consists of a window of 9 frequency channels and 11 time-frames. The number of nodes in the hidden layer of the MLPs is fixed to 100. Like in the first stage, they are trained for 100 epochs using the cross-entropy learning criterion.

4. RESULTS

4.1. Experimental setup

The proposed system is evaluated using the Aurora-4 dataset [23], which is based on the Wall Street Journal corpus [24]. The DNNs

[2] We define SNR span as the difference between the instantaneous SNRs corresponding to the desired target values of 0.95 and 0.05.

and the MLPs are trained using the noisy utterances from the multi-condition training set. These utterances were created by mixing speech with 6 noise types at SNRs ranging from 10 dB to 20 dB. Of the 2676 utterances in the set, 2100 sentences are used to train the system and the rest are used for cross validation and early stopping. The clean and noise signals comprising each mixture is used to set the desired target for training using Eq. 1. For evaluating performance, we use the reduced noisy test sets of the corpus. It consists of 166 clean utterances mixed with the same 6 noise types at SNRs ranging from 5 dB to 15 dB.

The ASR system is implemented using the HTK toolkit [25]. The recognition module consists of state tied, word-internal triphones modeled as 3-state HMMs. The observation probability of each state is modeled as a mixture of 16 diagonal Gaussians. The standard bigram language model and the CMU pronunciation dictionary are used during decoding. As features, we use 12th order MFCCs along with their delta and acceleration components. The features are mean and variance normalized at the utterance level to improve robustness. During testing, the noisy signals are filtered using the estimated IRM in the Mel spectral domain before cepstral transformation.

4.2. Evaluation results

4.2.1. *Instantaneous SNR estimation*

We first present the instantaneous SNR estimation performance of the proposed system. The proposed 2-stage system is compared with the following alternatives: 1) a 1-stage system that directly uses the output of the DNN without any smoothing, 2) a 1-stage system that directly estimates the IRM rather than the targets as defined by Eq. 1 (IRM-direct), 3) a system similar to the one proposed by Tchorz and Kollmeier [18] (TK-AMS). TK-AMS concatenates the AMS features calculated at the T-F unit level to obtain a frame-level feature (dimensionality: 15×26=390). A single DNN is then trained to simultaneously estimate the outputs corresponding to the 26 frequency channels. The architecture of the DNN is the same as used by the proposed system, except that the input and the output layers now consist of 390 and 26 nodes, respectively. The output of each of these systems is converted to decibels to evaluate performance. The ground truth instantaneous SNR values and the estimates are restricted to the range of -15 dB to 10 dB; any estimate out of this range is rounded to these boundary values.

The mean absolute error averaged across the 6 noise conditions is shown in Fig. 2. On average, the 1-stage system gives a mean error of 3.0 dB. Smoothing the output further improves the average error by 0.3 dB. Estimating the IRM directly worsens performance by around 1 dB compared to the 2-stage system, which shows the utility of using the sigmoid function to transform the SNRs. TK-AMS produces an average error of 3.7 dB.

It is interesting to note that the average error for every frequency channel is below 4 dB for the proposed 2-stage algorithm. It performs worst in babble noise and airport noise conditions, where the mean error across all channels is around 3 dB. As expected, it performs the best in the relatively stationary car noise conditions with an average error of 2.3 dB. It can be observed from the figure that the performance of all algorithms drops at higher frequency channels. This is expected since the high frequency region contains more unvoiced speech, which has noise-like characteristics, making it difficult to distinguish it from actual noise.

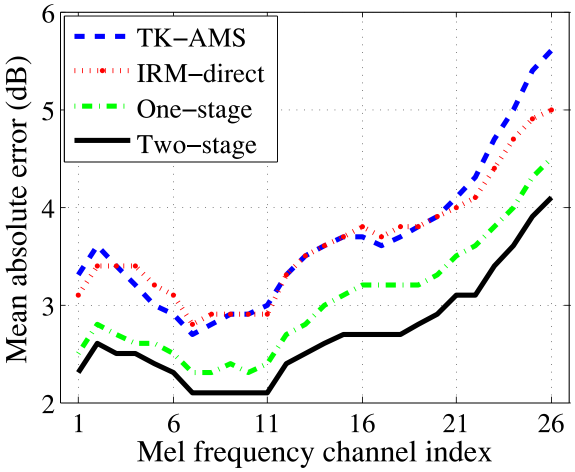

Fig. 2. Instantaneous SNR estimation performance in the Mel spectral domain. 26 frequency channels span frequencies from 50 Hz to 7 kHz.

4.2.2. *ASR performance*

We use both the clean and multi-condition (MC) training sets to train two ASR systems. In clean conditions they produce a word error rate (WER) of 8% and 10.4%, respectively. The performance of the tested feature enhancement algorithms are shown in Table 1. The baseline performance corresponds to recognizing noisy speech directly without any enhancement. This results in an average WER of 29% when using models trained in clean conditions, and 19.3% using MC training.

Apart from the systems described before, we also present results obtained using an IBM estimation algorithm (IBM-direct). The IBM-direct system uses binary targets, instead of ratio targets as used by the proposed algorithm, during training. The binary targets are obtained by applying a threshold to the instantaneous SNR at -6 dB. It uses DNNs trained similarly to the proposed 1-stage system (i.e., without any smoothing). The IBM-direct system is most similar to the one proposed in [21], which has been shown to perform well for speech separation. The direct masking approach is used to perform feature enhancement when the estimated IBM is used.

As can be seen, using models trained in clean conditions, the proposed 2-stage system obtains an average WER of 17.9%, 0.7 percentage points better than the 1-stage system and 3.6 percentage points better than IBM-direct. Clearly, estimating the IRM seems more appropriate for the task of ASR. Both 1-stage and 2-stage systems outperform IRM-direct and TK-AMS. It is worth emphasizing that the 2-stage system obtains a large improvement of 11.1 percentage points when compared to the noisy baseline. The difference in performance is not as dramatic when the ASR models are trained using the MC set. The 2-stage system obtains an improvement of 0.5 percentage points over the 1-stage system. The remaining systems obtain similar WERs on average. Compared to the noisy baseline, the 2-stage system obtains an improvement of 2.8 percentage points.

5. CONCLUSIONS

We have proposed a feature enhancement algorithm for improving noise robustness of ASR systems. The algorithm estimates a smoothed ideal ratio mask in the Mel spectrogram domain using

Table 1. Word error rates on the noisy subset of the Aurora4 corpus. The proposed systems are denoted as One-stage and Two-stage. RI stands for relative improvement with respect the noisy baseline.

System	Test set							
	Car	**Babble**	**Restaurant**	**Street**	**Airport**	**Train**	**Average**	**RI**
Clean Training								
Noisy	16.7	29.8	31.9	29.9	29.4	36.3	29.0	0%
IRM-direct	12.0	19.5	23.8	20.7	20.8	20.7	19.6	32.5%
TK-AMS	11.6	21.1	21.0	20.2	20.3	24.2	19.7	31.9%
IBM-direct	13.4	21.8	25.1	22.5	21.0	25.1	21.5	25.9%
One-stage	11.0	19.1	22.6	19.2	19.5	20.0	18.6	35.9%
Two-stage	10.7	19.2	21.7	18.0	18.9	18.7	17.9	38.4%
Multi-condition Training								
Noisy	12.9	17.4	23.9	20.0	18.8	22.7	19.3	0%
IRM-direct	11.8	18.5	21.7	17.1	19.4	18.1	17.8	7.9%
TK-AMS	11.7	17.5	19.5	17.0	18.4	18.9	17.2	11.0%
IBM-direct	12.6	17.2	20.5	17.7	17.9	18.6	17.4	9.6%
One-stage	11.0	18.0	20.2	16.8	18.3	17.6	17.0	11.9%
Two-stage	11.4	16.8	19.6	16.4	18.2	16.8	16.5	14.3%

deep neural networks, which is used to filter out noise before cepstral transformation. Large improvements were obtained on the Aurora-4 robust ASR task using the proposed system. It is also observed that better ASR performance is obtained using the estimated ratio mask compared to the estimated binary mask. We note that the noise types used in the test set are seen during training. Therefore, an interesting issue for future study is how well the system generalizes to unseen conditions.

Acknowledgements. We thank Yuxuan Wang for very helpful discussions and for providing the DNN training scripts.

6. REFERENCES

[1] T. Virtanen, B. Raj, and R. Singh, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*, John Wiley & Sons, West Sussex, UK, 2012.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578 –589, 1994.

[3] ETSI, ES 202 050 V1.1.4, "Speech processing transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," 2005.

[4] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech and language*, vol. 12, no. 2, pp. 75–98, 1998.

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

[6] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Computer, Speech, and Language*, vol. 23, pp. 389–405, 2009.

[7] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[8] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, Florida, 2007.

[9] J. Droppo, L. Deng, and A. Acero, "Improvements to VTS feature enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012, pp. 4677–4680.

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

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

[12] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and D. L. Wang, "Nothing doing: Re-evaluating missing feature ASR," Tech. Rep. OSU-CISRC-7/11-TR21, Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA, 2011, Available: ftp://ftp.cse.ohio-state.edu/pub/tech-report/2011/.

[13] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[14] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifer for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[15] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, pp. 1486–1501, 2006.

[16] J. Barker, L. Josifovski, M. P. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 373–376.

[17] J. van Hout and A. Alwan, "A novel approach to soft-mask estimation and log-spectral enhancement for robust speech recognition," in *Proceedings of the IEEE International Conference*

on Acoustics, Speech, and Signal Processing, 2012, pp. 4105–4108.

[18] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 11, pp. 184–192, 2003.

[19] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *Journal of Acoustical Society of America*, vol. 125, pp. 2336–2347, 2009.

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

[21] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, in press.

[22] G.E. Hinton, S. Osindero, and Y.W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[23] N. Parihar and J. Picone, "Analysis of the Aurora large vocabulary evalutions," in *Proceedings of the European Conference on Speech Communication and Technology*, 2003, pp. 337–340.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993 [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html.

[25] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Publishing Department, 2002, [Online]. Available: http://htk.eng.cam.ac.uk.