ROBUST SPEAKER RECOGNITION BASED ON DNN/I-VECTORS AND SPEECH SEPARATION

Jorge Chang[1] and DeLiang Wang[1,2]

Department of Computer Science and Engineering[1]
& Center for Cognitive and Brain Sciences[2]
The Ohio State University, Columbus OH, USA

changcheng.1@osu.edu, dwang@cse.ohio-state.edu

ABSTRACT

Recent research shows that the i-vector framework for speaker recognition can significantly benefit from phonetic information. A common approach is to use a deep neural network (DNN) trained for automatic speech recognition to generate a universal background model (UBM). Studies in this area have been done in relatively clean conditions. However, strong background noise is known to severely reduce speaker recognition performance. This study investigates a phonetically-aware i-vector system in noisy conditions. We propose a front-end to tackle the noise problem by performing speech separation and examine its performance for both verification and identification tasks. The proposed separation system trains a DNN to estimate the ideal ratio mask of the noisy speech. The separated speech is then used to extract enhanced features for the i-vector framework. We compare the proposed system against a multi-condition trained baseline and a traditional GMM-UBM i-vector system. Our proposed system provides an absolute average improvement of 8% in identification accuracy and 1.2% in equal error rate.

Index Terms— Robust speaker recognition, deep neural networks, i-vector, Speech Separation, time-frequency masking.

1. INTRODUCTION

Automatic speaker recognition is the task of recognizing the identity of a speaker from the speech signal. The task can be divided into speaker verification (SV) and speaker identification (SID). The objective of speaker verification is to verify an identity claim using the voice of the subject. In speaker identification, the goal is to provide the identity of the subject. Speaker recognition has many real world applications, including user authentication, access control, and assistance to speech separation and recognition.

Introduced by Dehak *et al.* [2], the i-vector framework is the dominating approach in speaker recognition research. This approach extends from a joint factor analysis which models speaker and channel subspaces separately [6]. In contrast, the i-vector approach models speaker and channel variations together. This leads to increased robustness to channel variations and other signal distortions.

The recent success of deep neural networks (DNNs) for automatic speech recognition (ASR) [5] motivated the application of DNNs to speaker recognition. Some methods train DNNs to directly distinguish speakers and have been primarily used in text-dependent speaker recognition (e.g. [8, 21]). Others use DNNs trained for a different task to extract information to aid subsequent speaker recognition [18]. One such approach uses a DNN trained for ASR as a universal background model (UBM) for an i-vector system, enhancing its ability to capture pronunciation patterns. This results in significant improvements over the traditional Gaussian mixture model (GMM) UBM [7, 19]. However, these studies focus on relatively clean conditions where the voice of the target speaker is not much interfered.

This paper investigates the performance of a phonetically-aware i-vector system in conditions where the speech is corrupted by strong additive noise. A supervised speech separation algorithm is proposed to remove or attenuate background noise. Specifically, a DNN is trained for mask estimation, and the DNN generated mask is used to extract enhanced input features for a DNN/i-vector framework.

The rest of the paper is organized as follows. In Section 2 we describe the proposed system. Section 3 presents the dataset and evaluation metrics. We then describe the results of our experiments and comparisons in Section 4. Finally, concluding remarks are given in Section 5.

2. SYSTEM DESCRIPTION

This section provides the background and describes each element of the proposed system. A diagram of the proposed system is shown in Figure 1.

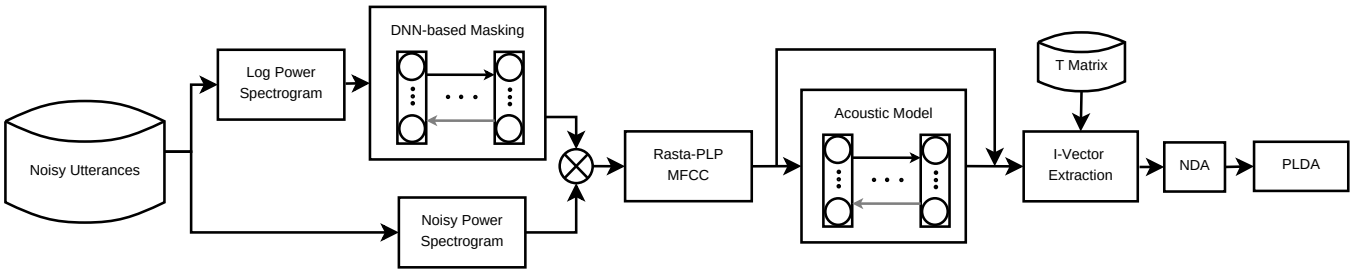

Fig. 1: Schematic diagram of the proposed system. See text for explanations of acronyms.

2.1. Speech Separation

We perform speech separation by estimating the ideal ratio mask (IRM) in a supervised fashion. The IRM is defined as the ratio between the energy of the clean and noisy speech at each time-frequency (T-F) unit [22]. As speech and noise can be assumed to be uncorrelated, the energy of the mixture becomes the sum of the speech energy and noise energy. The IRM can be defined in different T-F domains. We define the IRM in the power spectrogram domain:

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

where S and N denote the power spectrogram of the speech and noise respectively. The resulting IRM is closely related to the Wiener filter.

We employ a DNN as the discriminative learning machine to estimate the IRM. The network inputs consist of the log-compressed power spectrogram frames. For the current frame, we take the preceding and succeeding 10 frames, resulting in a 2,709 dimensional (129×21) feature vector. For the training target, we use the IRM corresponding to the central frame of the input vector along with the preceding and succeeding two frames, resulting in a 645 dimensional (129×5) feature vector. Taking a context window of the current input frame allows the network to leverage the temporal dynamics of speech. For the output, a context window provides several estimates of the IRM at each frame, whose average produces a better estimate [22]. At the test stage, the estimated IRM is pointwise multiplied with the noisy power spectrogram to produce an enhanced power spectrogram which is used for feature extraction in the i-vector system.

The DNN used in our experiments consists of four hidden layers. Each hidden layer contains 1024 rectified linear units (ReLUs). The network is trained to minimize the mean square error using stochastic gradient descent and Adagrad [3]. Dropout regularization is set to 0.2. A momentum term is included and set to 0.2 for the first 30 epochs and 0.8 afterwards. The DNN is trained for a total of 150 epochs.

2.2. I-Vector Algorithm

The i-vector algorithm projects input features into a low-dimensional space. It assumes that a speaker- and session-dependent supervector M can be modeled as:

$$M = m + Tw \tag{2}$$

where m represents the speaker- and channel-independent component, T is a matrix of a low rank referred as the total variability matrix, and w is a random vector known as the i-vector [2]. To learn the bases for the total variability subspace, Baum-Welch statistics are computed from a UBM.

After i-vectors are extracted, linear discriminant analysis (LDA) is typically used to further reduce the dimensionality of the i-vectors. However, nonparametric discriminant analysis (NDA) has been shown to be more effective [19, 9]. This is attributed to the Gaussian assumption in LDA. It is known that this assumption may not hold, especially in the presence of noise and channel variations. In NDA, local sample averages are computed based of the k-nearest neighbors of a sample to replace the global information about each class [19].

We implement our speaker recognition systems using the MSR Identity Toolbox [20]. The DNN acoustic model consists of 7 fully connected hidden layers each with 2048 ReLUs. The DNN is trained to estimate the clean posterior probability of 4096 senones. Target senone states are obtained using a GMM-HMM system built with the Kaldi toolkit [17]. The network is trained to minimize the cross-entropy loss function. All other settings are identical to the ones in the previous section. For the GMM-UBM, the number of Gaussian components is set to 2048 which showed the best results among different configurations. We use a 400 dimensional total variability matrix trained with the expectation maximization algorithm to extract the i-vectors. Afterwards, we apply NDA with $k = 8$ to reduce the dimensionality of each vector to 200 before using them to train a probabilistic linear discriminant analysis (PLDA) model.

For our features, we use 19 mel-frequency cepstral coefficients (MFCCs), 13 relative spectral filtered perceptual linear predictive cepstral coefficients (RASTA-PLP) and the log energy of each frame. Delta and double delta coefficients are included for a total of 99 (33×3) dimensions per frame.

SNR	-5 dB		0 dB		5 dB		10 dB		Average	
System	ACC	EER	ACC	EER	ACC	EER	ACC	EER	ACC	EER
GMM/i-vector	39.05	11.26	62.50	7.37	68.69	6.06	69.59	5.93	59.96	7.65
GMM/i-vector + Masking	43.04	11.14	68.94	6.18	78.22	5.00	79.51	4.77	67.43	6.77
DNN/i-vector	42.27	10.87	66.24	5.97	73.20	5.59	74.10	5.74	63.95	7.04
DNN/i-vector + Masking	42.01	9.79	69.59	5.83	78.61	5.00	80.54	4.35	67.69	6.24

Table 1: EER (%) and SID accuracy (%) under matched SNR conditions with SSN.

SNR	-5 dB		0 dB		5 dB		10 dB		Average	
System	ACC	EER	ACC	EER	ACC	EER	ACC	EER	ACC	EER
GMM/i-vector	31.19	13.23	55.28	9.01	65.08	7.00	66.88	5.80	54.61	8.76
GMM/i-vector + Masking	38.02	12.14	61.21	7.47	75.90	5.34	76.42	5.03	62.89	7.50
DNN/i-vector	35.95	11.76	61.08	6.44	69.59	5.84	70.62	5.67	59.31	7.43
DNN/i-vector + Masking	38.27	11.38	64.43	6.19	76.93	4.60	78.74	4.24	64.59	6.60

Table 2: EER (%) and SID accuracy (%) under matched SNR conditions with babble noise.

3. EXPERIMENTAL SETUP

3.1. Dataset

Acoustic models are trained using the Switchboard-1 corpus [4]. Speaker recognition experiments are conducted on female speakers from the NIST SRE 2006 [16] and 2008 [15] dataset ($8conv$ condition). We use SRE 2006 for development and SRE 2008 for enrollment and testing. The speech separation DNN is trained on unused data from NIST SRE 2006 and Switchboard-1.

A total of 402 and 395 speakers are used from the NIST SRE 2006 and 2008 dataset, respectively. For each target speaker eight two-channel telephone conversations are provided. Each of these conversations contains about two minutes of speech from the target speaker. For each utterance we remove the non-target speaker and pass it through a voice activity detector to remove large chunks of silence. The utterance is then divided into 15 second pieces, ensuring that the total energy of each piece is above an empirical threshold. Speakers without sufficient quality data (i.e. less than two minutes of data across all 8 utterances) are discarded. Utterances are then mixed with babble or speech-shaped noise (SSN) at signal-to-noise ratios (SNR) ranging from -5 dB to 12 dB to produce the noisy utterances. Each noise is four minutes long and is divided into three pieces. We assign one piece for speech separation, and two pieces for i-vector training and testing.

3.2. Evaluation Metrics

Experiments are evaluated using equal error rate (EER) for the verification task and identification accuracy (ACC) for the identification task. The EER represents the value at which the false positive rate equals the false negative rate when comparing each testing sample against all speakers in the test set. The identification accuracy is the percentage of correct identifications among all the test trials. Both of these metrics are commonly used for evaluating speaker recognition systems.

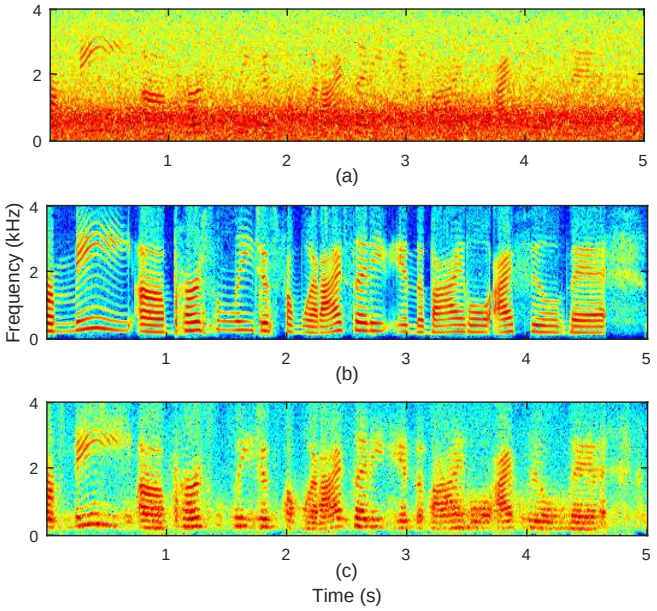

Fig. 2: Speech separation illustration. (a) Power spectrogram of a speech utterance mixed with SSN at -5 dB SNR, (b) Power spectrogram of the original speech signal, and (c) Power spectrogram of the separated speech signal.

4. RESULTS AND DISCUSSION

First, we show an example of a speech separation result in Figure 2. The speech signal is taken randomly from our test set. We can see that the power spectrogram of the separated

SNR	-3 dB		3 dB		8 dB		12 dB		Average	
System	ACC	EER	ACC	EER	ACC	EER	ACC	EER	ACC	EER
GMM/i-vector	36.08	11.47	63.27	6.75	67.14	5.97	68.56	5.67	58.76	7.47
GMM/i-vector + Masking	53.61	9.02	75.39	5.41	79.38	4.90	79.61	4.64	72.00	5.99
DNN/i-vector	37.37	11.38	66.75	5.90	72.68	5.78	72.42	5.03	62.31	7.02
DNN/i-vector + Masking	59.28	7.47	76.03	4.60	79.77	4.12	80.28	4.07	73.84	5.07

Table 3: EER (%) and SID accuracy (%) under unmatched SNR conditions with SSN.

SNR	-3 dB		3 dB		8 dB		12 dB		Average	
System	ACC	EER	ACC	EER	ACC	EER	ACC	EER	ACC	EER
GMM/i-vector	38.27	11.14	61.34	7.86	66.88	6.23	67.40	5.86	58.47	7.77
GMM/i-vector + Masking	45.62	10.44	71.13	5.93	77.19	5.06	76.29	5.03	67.56	6.61
DNN/i-vector	44.21	9.28	66.19	6.22	70.67	5.67	70.67	5.59	62.94	6.69
DNN/i-vector + Masking	46.70	8.33	72.54	5.23	78.92	4.51	77.79	4.43	68.99	5.63

Table 4: EER (%) and SID accuracy (%) under unmatched SNR conditions with babble noise.

utterance recovers much of the structure of the clean signal.

We then evaluate a GMM/i-vector and a DNN/i-vector baseline in clean conditions. The GMM system achieves a 2.96% in EER and 92.91% in SID accuracy. The DNN system achieves 1.29% in EER and 96.65% in SID accuracy. The DNN system clearly outperforms the GMM system, which is consistent with previous studies [19, 7].

Table 1 shows the results of our experiments conducted in matched SNR conditions and with SSN. These SNRs are the ones used to train all the systems. Both DNN and GMM systems benefit consistently from masking, with higher improvements at higher SNRs. From the average improvements, we observe that the GMM system benefits more from speech separation with an average improvement of 7.9% in accuracy and 1.1% in EER in absolute terms. For the DNN system, the average improvement is 4.5% in accuracy and 0.8% in EER. This may be due to the fact that, with its better performance, the DNN/i-vector approach has less room to improve. This observation is supported by comparing the two baseline systems in clean conditions.

Table 2 shows the ACC and EER results in babble noise, a non-stationary noise. All systems perform worse compared to the previous experiments in the stationary noise of SSN. The average performance drop is 4% and 1.5% in accuracy and EER respectively, indicating the more challenging nature of babble noise. However, we also see that T-F masking provides higher improvements in these conditions.

We also conduct experiments in unmatched SNR conditions, where the average input SNR is 5 dB, the same as in the matched SNR conditions. The results for these experiments are presented in Tables 3 and 4. We see the same patterns of speaker recognition results from the previous experiments. All systems generalize well to unseen conditions. At 12 dB the performance appears to be lower than expected, underperforming 8 dB trials in some cases. This suggests that the systems do not perform well outside the training SNR range (10 dB to -5 dB). The rest of the results are consistent with the previous experiments.

5. CONCLUSIONS

In this study, we propose a front-end speech separation system for the i-vector framework to deal with utterances corrupted by background noise. This is done using a DNN to estimate the IRM. The separation front-end is applied to a standard GMM/i-vector system and a DNN/i-vector system. We show that the speech separation improves the performance of the baseline systems for both identification and verification tasks. The overall average improvement from using speech separation is 8% (absolute) in identification accuracy and 1.2% (absolute) in EER.

To our knowledge, this is the first study combining i-vector based speaker recognition and DNN based speech separation. Recent advances in speech separation suggest avenues to improve mask estimation [1, 24]. Studies in acoustic modeling also suggest alternative ways to perform speech separation with acoustic models [10, 23].

6. ACKNOWLEDGMENTS

The authors would like to thank Zhong-Qiu Wang and Jitong Chen for helpful discussions and Xiaojia Zhao for providing some of the tools used. This research was supported in part by an AFOSR grant (FA9550-12-1-0130) and the Ohio Supercomputer Center.

7. REFERENCES

[1] Chen J., Wang Y., Yoho S.E., Wang D.L. and Healy E.W., " Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *Journal of the Acoustical Society of America*, vol. 139, pp. 2604-2612, 2016.

[2] Dehak N., Kenny P.J., Dehak R., Dumouchel P. and Ouellet P., "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788-798, 2011.

[3] Duchi J., Hazan E. and Singer Y., "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121-2159, 2011.

[4] Godfrey J. and Holliman E., "Switchboard-1 Release 2 LDC97S62," *Philadelphia: Linguistic Data Consortium*, 1993.

[5] Hinton G., *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 82-97, 2012.

[6] Kenny P., "Joint factor analysis of speaker and session variability: Theory and Algorithms," *Tech Report CRIM-06/08-13*, 2005.

[7] Lei Y., Scheffer N., Ferrer L. and McLaren M., "A novel scheme for speaker recognition using a phonetically-aware deep neural network," *Proceedings of ICASSP*, pp. 1714-1718, 2014.

[8] Li L., Lin Y., Zhang Z. and Wang D., "Improved deep speaker feature learning for text-dependent speaker recognition," *Proceedings of APSIPA Annual Summit and Conference*, pp. 426-429, 2015.

[9] Li N. and Mak M.W., "SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1648-1670, 2015.

[10] Narayanan A. and Wang D.L., " Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 92-101, 2015.

[15] NIST Multimodal Information Group, "2008 NIST Speaker Recognition Evaluation Training Set Part 1 LDC2011S05," *Philadelphia: Linguistic Data Consortium*, 2011.

[16] NIST Multimodal Information Group, "2006 NIST Speaker Recognition Evaluation Training Set LDC2011S09," *Philadelphia: Linguistic Data Consortium*, 2011.

[17] Povey D., Ghoshal A. and Boulianne G., "The Kaldi speech recognition toolkit," *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[18] Richardson F., Reynolds D.A. and Dehak N., "Deep Neural Network Approaches to Speaker and Language Recognition," *IEEE Signal Processing Letters*, vol. 22, pp. 1671-1676, 2015.

[19] Sadjadi S.O., Ganapathy S. and Zhu W., "The IBM 2016 speaker recognition system," *Proceedings of Speaker and Language Recognition Workshop (Odyssey)*, pp.174-180, 2016.

[20] Sadjadi S.O., Slaney M. and Heck L., "MSR Identity Toolbox v1.0: A MATLAB toolbox for speaker-recognition Research," *Speech and Language Processing Technical Committee Newsletter*, 2013.

[21] Variani E., Lei X., McDermott E., Lopez Moreno I. and Gonzales-Dominguez J., "Deep neural networks for small footprint text-dependent speaker verification," *Proceedings of ICASSP*, pp. 4080-4084, 2014.

[22] Wang Y., Narayanan A. and Wang D.L., "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849-1858, 2014.

[23] Wang Z.Q. and Wang D.L., " A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 796-806, 2014.

[24] Williamson D.S., Wang Y. and Wang D.L., "Complex ratio masking for monaural speech separation," *EEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 483-492, 2016.

[25] Zhao X., Wang Y. and Wang D.L., "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 836-845, 2014.