Text-available speaker recognition system for forensic applications

Chengzhu Yu,Chunlei Zhang, Finnian Kelly, Abhijeet Sangwan, John H. L. Hansen

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX, U.S.A.
{chengzhu.yu, john.hansen}@utdallas.edu

Abstract

This paper examines a text-available speaker recognition approach targeting scenarios where the transcripts of test utterances are either available or obtainable through manual transcription. Forensic speaker recognition is one of such applications where the human supervision can be expected. In our study, we extend an existing Deep Neural Network (DNN) i-vector-based speaker recognition system to effectively incorporate text information associated with test utterances. We first show experimentally that speaker recognition performance drops significantly if the DNN output posteriors are directly replaced with their target *senone*, obtained from force alignment. The cause of such performance drops can be attributed to the fact that forced alignment selects only the single most probable *senone* as their output, which is not desirable in a current speaker recognition framework. To resolve this problem, we propose a posterior mapping approach where the relationship between forced aligned *senonoes* and its corresponding DNN posteriors are modeled. By replacing DNN output posteriors with *senone* mapped posteriors, a robust text-available speaker recognition system can be obtained in mismatched environments. Experiments using the proposed approach are performed on the Aurora-4 dataset.

Index Terms: speaker recognition, forensic speaker recognition

1. Introduction

Research on speaker recognition has focused either on text-dependent or text-independent scenarios [1–3]. Text-dependent speaker recognition assumes that the same speech content is used for enrollment and recognition. On the other hand, text-independent speaker recognition does not have any constraint on the speech content. Moreover, most of the text-independent speaker recognition system assume that the content of speech utterances is unknown. However, in many speaker recognition scenarios, the text of speech utterances for both enrollment and recognition could be obtained through manual transcription.

Forensic speaker recognition [4] is one such application; the nature of forensic casework means that there is always supervision of the automatic system by the forensic expert. As laid out in recent European guidelines [5], before any speech content is passed to an automatic system, it should first be manually evaluated by the forensic expert. The ‘quality of the speech content is assessed in terms several factors, including its duration, the vocal effort and emotional state of the speaker, along with environmental noise or channel effects. If these quality factors are satisfactory, the speaker is broadly profiled in terms of gender, age and accent, informing the choice of a suitable speaker population for calibration/normalization within the automatic system. At this point, a manual transcription of the speech could be made, and potentially passed to the automatic system along with the speech recordings. This transcription of the evidential recording would form part of the experts case report.

One of the major challenges in speaker recognition is the mismatch between the speech samples under comparisons [6–9]. In this paper, we investigate the use of text information to improve the robustness of a speaker recognition system in mismatched conditions. Previous studies has investigated text-available speaker recognition problems, with most of them focused on text prompt speaker recognition [10–13]. However, to the best of our knowledge, no study has yet been able to effectively incorporate text information to improve the speaker recognition system from the perspective of speaker modeling, particularly in a state-of-the art speaker recognition system based on i-vector extraction and PLDA modeling.

The main reasons that text information has not been effectively used in i-vector based speaker recognition systems is the lack of direct correspondence between the universal background model (UBM) and a speech transcript. The recently proposed DNN-based i-vector extraction [14] has the potential to overcome such limitation, as the traditional UBM is replaced with tied-state triphones (*senones*), that are strongly correlated with the content of the underlying speech utterances. In fact, the target output of each speech frame used for training DNN system comes directly from the *senone* labels obtained from forced alignment of speech transcripts. However, the experiments in our study indicate that by directly replacing DNN predictions with its ground truth target *senone* significantly decreases the performance. This is mainly due to the fact that the use of hard alignment, e.g.,

force alignment, introduces inaccuracies that are better avoided when using i-vector based approaches.

To resolve the hard alignment problem of using *senone* forced alignment in speaker recognition, we propose to model the relationship between target *senone* and its corresponding DNN prediction probabilities in training data. This is achieved by averaging the DNN prediction probabilities from frames that are aligned with each *senone* target.

At the test time, instead of using the posteriors obtained from DNN prediction, the posteriors mapped from corresponding *senones* are used for both total variability matrix training and i-vector extraction. The advantage of using mapped posteriors from force aligned *senones* is most significant in mismatched environments where the DNN prediction becomes less reliable. The force alignment is expected to be more robust compare to DNN prediction in noisy environments, as a strong prior information is provided by means of speech transcript.

In Sec. 2, we present a short overview of the DNN based i-Vector extraction. Sec. 3 contains the description of proposed method. In Sec. 4 and 5, we present results to show the effectiveness of proposed framework.

2. i-vector extraction

2.1. UBM based i-vector extraction

In i-vector extraction framework, speaker and channel dependent Gaussian mixture model (GMM) supervector is modeled as belows:

$$M = m + Tw, \tag{1}$$

where m is the supervector generated from the UBM means, T is the total variability matrix formed by the basis of reduced total variability space, and w is the factor loading also known as i-vectors.

The total variability matrix T is estimated by using expectation maximization (EM) method as described in [15]. After total variability matrix training, the i-vector of each speech utterance can be represented using Baum-Welch zeroth (N_s) and centralized first (F_s) order statistics:

$$w_s^* = (T' N_s \Sigma^{-1} T + I)^{-1} T \Sigma^{-1} F_s, \tag{2}$$

where Σ is the covariance matrix obtained from UBM model and I is the identity matrix. Here, N_s and F_s are expressed as

$$N_s = \begin{bmatrix} N_s^{C=1} & 0 & 0 & 0 \\ 0 & N_s^{C=2} & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & N_s^{C=c} \end{bmatrix}, \tag{3}$$

$$F_s = \begin{bmatrix} F_s^{C=1} & 0 & 0 & 0 \\ 0 & F_s^{C=2} & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & F_s^{C=c} \end{bmatrix}, \tag{4}$$

where

$$N_s^{C=c} = \sum_t P(c|X_t, \theta_{UBM}), \tag{5}$$

$$F_s^{C=c} = \sum_t P(c|X_t, \theta_{UBM})(X_t - \mu_c). \tag{6}$$

and c is the index of UBM mixture component, X_t is acoustic feature at time t, μ_c is the mean of cth Gaussian component.

2.2. DNN based i-vector extraction

In DNN based i-vector extraction approach [14], the UBM is replaced with stacked *senones* and the posterior probabilities of each speech belongs to individual *senones* are obtained with DNN predictions. During the training of DNN model, the target *senones* of each frame are generated by force aligning each acoustic frame with given speech transcript.

2.3. Force alignment based i-vector extraction

As the prediction target of DNN for building i-vector system is obtained from force alignment, the straight forward way of introducing text information in speaker recognition is to replace the predicted DNN posteriors with its ground truth *senone* target obtained from transcripts. The posterior vector of each speech frames obtained from force alignment is a vector of zeros with only the senone that aligned with corresponding speech frames being one. However, our experiment results show that the direct replacement of DNN posteriors with force aligned *senone* estimation, decrease the performance dramatically (Table. 1).

2.4. DNN posterior weighting

The reason of such performance loss when using ground truth *senone* target is that the output of force alignment assign only single *senone* to each frame. The limitations of such hard alignment could be manifold including the sparseness of speech data in accordance with certain *senones*.

To overcome such limitation, we apply the fusion of DNN posterior with forced aligned *senone* prediction. That is, the DNN posterior probability associated with aligned *senones* are forced to increase as shown below.

$$\begin{aligned} p'(k|x_t) &= \alpha \times p(k|x_t) + \alpha \times 1, \text{ if } s_t = k \\ p'(k|x_t) &= \alpha \times p(k|x_t) + \alpha \times 0, \text{ if } s_t \neq k \end{aligned} \tag{7}$$

where k is *senone* id, $p(k|x_t)$ indicates the DNN prediction probability associated with kth *senone*. The α

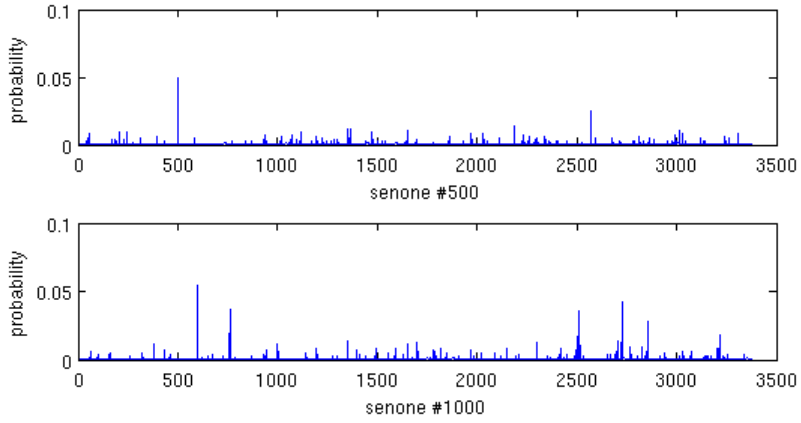

Figure 1: *The mapped posteior vector for senone-500 and senone-1000.*

is weighting factor that control the contributions of each part. In our experiment, a good performance is obtained by setting α as 0.9.

2.5. Posterior mapping

While the DNN posterior weighting in Sec. 2.4 could effectively incorporate alignment information from transcript in a softer way than directly using force alignment, it is not theoretically well motivated and involves a use of hyper-parameter for fusing the posteriors from two different sources. In this section, we propose another approach, which models the correlation between force alignment and DNN prediction in the training data. The objective of this approach is to find the maximum likelihood DNN posterior vectors given force alignment. The posteriors obtained with this approach is more robust to the ones obtained from DNN prediction, as the effect of noise or mismatch environment on force alignment is much less than DNN prediction due to supervised information from transcription.

Specifically, we model the correlation between force alignment and DNN posterior probabilities using the posterior probabilities obtained from DNN training data. As during training, there is an one to one mapping between force alignment and DNN posteriors for each speech frame. Note that the data used for training this relationship has no overlap with the ones used for composing speaker verification trials. We model the correlation between DNN posteriors and their targets *senone* which obtained from force alignment as follow:

$$M_k = \frac{\sum_{s_t=k} \boldsymbol{p}_t}{T} \tag{8}$$

where M_k is the average of DNN posterior probability associated with target *senone* k, $\boldsymbol{p}_t$ is the DNN posterior prediction vector as

$$\boldsymbol{p}_t = [p(s_1|x_t), p(s_2|x_t), ..., p(s_K|x_t)], \tag{9}$$

T is the total frame number,. After obtain M_k for each *senone* using training, we replace $\boldsymbol{p}_t$ with M_k, if the speech frame x_t is aligned with k-th *senone*. We perform such posterior mapping during both total variability matrix training and i-vector extraction states.

Fig. 1 is an example of posterior mapping obtained from $500th$ and $1000th$ *senone* respectively. The upper plot of Fig. 1 is the average posterior prediction on the training data when target *Senone* is 500. As expected, the average posterior probability on *Senone* 500 is much higher than the other ones. However, it is interesting to notice in the lower plot of Fig. 1 that the highest posterior probability when target *Senone* is 1000 is actually other *senones*. This indicates that the prediction of *Senone* 2000 is easily confused by other *Senones*.

3. Experiments

3.1. System setup

As the standard speaker evaluation corpus such as NIST SRE does not include speech transcripts, the proposed text based speaker recognition system is evaluated on Aurora-4 database. The Aurora-4 is noisy version of Wall Street Journal (WSJ0) corpus. The multi-condition training set including 7137 utterances from 83 speakers. Half of the training utterances are obtained from the primary Sennheiser microphone. The other half are recordings from different secondary microphones. Part of those utterances are clean speech without noise and the other part are consists of corrupted utterances with six different noises (street traffic, car, train station, babble, airport, restaurant) at 10-20 dB SNR.

The UBM based i-vector extraction system are trained on multi-condition training set. The 2048 mixture of UBM and 400 dimension total variability matrix is trained on MFCC features of 39 dimension (13+△+△△). For backend verification both the cosine distance similarity (CDS) measure and probabilistic linear discriminant

Table 1: Experiment results on trials with utterances including both noisy and clean one.

	CDS	PLDA
I-vector (UBM)	19.48	8.15
I-vector (DNN)	17.87	8.84
I-vector (FA)	29.93	12.29
I-vector (DNN+weight)	16.30	8.00
I-vector (DNN+mapping)	15.91	6.67

analysis (PLDA) are used for evaluation [16, 17].

For bulding DNN based i-vector extraction system [18], GMM-HMM models with 3024 distinct tied-state triphones are trained using MFCC features along with their linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT). The alignment obtained from GMM-HMM system is then used for training DNN-HMM system. For the DNN-HMM systems, we first generatively pretrain the DNN with 7 layers of stacked RBM with 2048 hidden nodes in each layer. The DNN-HMM system was trained with 40 dimensional log Mel filterbank (FBANK) features. We use 256 minibatch and 0.008 as the start learning rate. After each epoch of training, the learning rate is reduced by half when the improvements in development set are less than 0.5%.

3.2. Results

The evaluation is performed on 2324 utterances from 8 unique speakers. A total of 806433 trials are created, including 98518 target trials. The experiments results are shown for both noisy trials in Table. 1 and clean trials in Table. 2. The results indicate that DNN based i-vector extraction does not show much advantages over UBM based i-vector extraction due to the limited amount of data used for training DNN. It can also be observed that the direct use of force alignment (FA), decrease the performance significantly. This indicates the importance of soft alignment in i-vector extraction. On the other hand, the proposed approaches based on DNN posterior weighting and DNN posterior mapping consistently outperforms both UBM and DNN based i-vector systems in noisy conditions. The relative improvement is higher using PLDA modeling approach. This can be explained by the fact that the utterances used for training PLDA is also noisy and therefore benefit more from proposed approaches. The best performing system is based on DNN posterior mapping in both noisy and clean evaluation setup.

Table 2: Experiment results on trials from clean utterances only.

	CDS	PLDA
I-vector (UBM)	5.39	2.05
I-vector (DNN)	6.72	2.45
I-vector (FA)	14.47	4.41
I-vector (DNN+weight)	5.66	1.79
I-vector (DNN+mapping)	7.11	1.27

4. Conclusion

In this study, we investigate the use of text information for improving the robustness of speaker recognition system. The proposed approached could potentially be beneficial for forensic speaker recognition where the human supervision can be expected. We evaluated the proposed systems using Aurora-4 database which has been widely used in the area of robust speech recognition. The experimental results indicate that proposed text available i-vector extraction framework consistently outperform conventional UBM and DNN based i-vector system. While the proposed approaches based on DNN-posterior weighting and mapping could effectively introduce text information into speaker modeling, further studies are needed to fully exploit the text knowledge.

5. Acknowledgments

This research was supported by National Science Foundation (NSF) under Grant 1219130.

6. References

[1] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *Signal Processing Magazine, IEEE*, vol. 32, no. 6, pp. 74–99, 2015.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[4] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. F. Bonastre, and D. Matrouf, "Forensic speaker recognition," Institute of Electrical and Electronics Engineers, 2009.

[5] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*, Frankfurt: Verlag f ür Polizeiwissenschaft, 2015.

[6] Yun Lei, Lukáš Burget, and Nicolas Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Proc. of ICASSP*, 2013, pp. 6788–6791.

[7] C. Yu, G. Liu, S. Hahm, and J. H. L. Hansen, "Uncertainty Propagation in Front End Factor Analysis For Noise Robust Speaker Recognition," *Proc. ICASSP, Florence, Italy*, 2014.

[8] C. Yu, G. Liu, and J. H. L. Hansen, "Acoustic Feature Transformation using UBM-based LDA for Speaker Recogntion," *Proc. Interspeech, Singapore*, 2014.

[9] Chunlei Zhang, Gang Liu, Chengzhu Yu, and John HL Hansen, "I-vector based physical task stress detection with different fusion strategies," *ISCA INTERSPEECH*, 2015.

[10] C. Che, Q. Lin, and D. Yuk, "An hmm approach to text-prompted speaker verification," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. IEEE, 1996, vol. 2, pp. 673–676.

[11] S. Nakagawa, W. Zhang, and M. Takahashi, "Text-independent/text-prompted speaker recognition by combining speaker-specific gmm with speaker adapted syllable-based hmm," *IEICE transactions on information and systems*, vol. 89, no. 3, pp. 1058–1065, 2006.

[12] F. S. Richardson and J. P. Campbell, "Transcript-dependent speaker recognition using mixer 1 and 2," *Target*, vol. 245, pp. 295–2, 2010.

[13] J. Lindberg and H. Melin, "Text-prompted versus sound-prompted passwords in speaker verification systems," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[14] Y. Lei, L. Ferrer, M. McLaren, et al., "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[15] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[16] Patrick Kenny, "Bayesian speaker verification with heavy tailed priors," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[17] Gang Liu, Chengzhu Yu, Abhinav Misra, Navid Shokouhi, and John HL Hansen, "Investigating state-of-the-art speaker verification in the case of unlabeled development data," in *Proc. Odyssey speaker and language recognition workshop, Joensuu, Finland*, 2014.

[18] C. Yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani, and J. H. L. Hansen, "Robust i-vector extraction for neural network adaptation in noisy environment," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.