DEEP NEURAL NETWORK BASED POSTERIORS FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Subhadeep Dey[1,2]*, Srikanth Madikeri*[1]*, Marc Ferras*[1] *and Petr Motlicek*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{subhadeep.dey, srikanth.madikeri, marc.ferras, petr.motlicek}@idiap.ch

ABSTRACT

The i-vector and Joint Factor Analysis (JFA) systems for text-dependent speaker verification use sufficient statistics computed from a speech utterance to estimate speaker models. These statistics average the acoustic information over the utterance thereby losing all the sequence information. In this paper, we study explicit content matching using Dynamic Time Warping (DTW) and present the best achievable error rates for speaker-dependent and speaker-independent content matching. For this purpose, a Deep Neural Network/Hidden Markov Model Automatic Speech Recognition (DNN/HMM ASR) system is used to extract content-related posterior probabilities. This approach outperforms systems using Gaussian mixture model posteriors by at least 50% Equal Error Rate (EER) on the RSR2015 in content mismatch trials. DNN posteriors are also used in i-vector and JFA systems, obtaining EERs as low as 0.02%.

Index Terms— Text-dependent speaker verification, DNN posterior, Dynamic Time Warping

1. INTRODUCTION

Text-dependent speaker verification aims at recognizing a person by matching voice characteristics and the message being spoken. As opposed to text-independent speaker recognition, where the message is unconstrained, both the speaker and the message must match to verify the speaker identity for text-dependent verification. Impostors can be divided into three categories, (i) the content does not match (ii) the speaker does not match (iii) neither the speaker or the content match.

Several approaches have been considered for text-dependent speaker recognition in the literature. A Hierarchical Multi-Layer Acoustic Model (HiLAM), using speaker-adapted Hidden Markov Model (HMM), was explored in [1, 2]. The state-of-the-art text-independent i-vector [3] approach has also been used, with the session variability term in Probabilistic Linear Discriminant Analysis (PLDA) jointly modeling speaker-content variability. In the same line, Joint Factor Analysis (JFA) [4] using speaker-content and session terms has been shown to perform well. I-vector and JFA approaches model content variability by pooling sufficient statistics, but do not consider any sequence information related to content. Alternatively, template matching techniques matching the speaker and the content have been used [5, 6]. The advantage of these methods is that factors such as speaking rate are normalized while scoring.

The model-based approaches mentioned earlier involve the computation of posterior probabilities from the components of a Gaussian Mixture Model (GMM), which is trained in an unsupervised manner. Recent research suggests that such posteriors can be replaced by posterior probabilities estimated using a DNN [7]. The DNN is trained discriminatively using frame labels obtained after forced alignment of a HMM/GMM acoustic model. Since transcripts are needed for forced alignment, the parameters of the DNN leading to state posterior estimates are trained in a supervised fashion, involving information unused in GMM training.

In this paper, we approach text-dependent speaker recognition using DNN posteriors in i-vectors and JFA frameworks. We hypothesize that using DNNs trained for Automatic Speech Recognition (ASR) systems result in linguistically meaningful posterior probabilities that allow to compare speaker characteristics in controlled contexts. We also explore DNN posteriors in DTW based systems showing that its efficiency for conditions detecting content mismatch is relevant.

The paper is organized as follows: Sections 2 and 3 describes the baseline system and the proposed DNN posterior approach respectively. Section 4 describes the experimental setup for evaluating the system and section 5 the results of the various system are discussed. Finally, we conclude in section 6.

2. BASELINE SYSTEM

A standard i-vector PLDA system is used as the baseline in our experiments [3, 8]. The i-vector system models a speech utterance as a low dimensional vector whose subspace is spanned by the columns of the total variability matrix, as

$$\mathbf{s} = \boldsymbol{\mu} + \mathbf{T}\mathbf{w} \,, \tag{1}$$

where $\mathbf{s}$ is the mean supervector, $\boldsymbol{\mu}$ is the mean supervector of a Universal Background Model (UBM). The matrix $\mathbf{T}$ is a low rank matrix projecting mean supervectors to obtain i-vectors $\mathbf{w}$, a low-dimensional representation of the audio recording. Undesirable channel effects can be removed from the i-vector using Linear Discriminant Analysis (LDA), whitening and length normalization, and PLDA.

To estimate the i-vector given a speech recording, we first estimate the zeroth and normalized first order statistics with respect to the UBM [9]. The zeroth order statistics are obtained by accumulating Gaussian component posteriors over all speech frames in an utterance. Similarly, the first order statistics accumulate the feature vectors per GMM component by weighting them with the corresponding posteriors.

Although i-vectors average out the time-varying content of an utterance, some studies suggest that the framework can still be relevant to text-dependent speaker recognition [1]. For short utterances,

as used in text-dependent speaker recognition, i-vector systems still provide speaker-discriminative scores.

Joint Factor Analysis (JFA) is used for text-dependent speaker recognition by explicitly modeling the content variability as a separate factor [10, 11]. Although, sequence information is still not modeled in this approach, recent developments suggest that state-of-the-art performance can be achieved. The JFA model,

$$\mathbf{s} = \boldsymbol{\mu} + \mathbf{D}\mathbf{z} + \mathbf{U}\mathbf{x}\,, \tag{2}$$

includes a diagonal term $\mathbf{D}\mathbf{z}$, with $\mathbf{D}$ being a diagonal matrix capturing the speaker variabilities, $\mathbf{z}$, the corresponding latent vector representing a speaker, $\mathbf{U}$, the eigenchannel matrix and $\mathbf{x}$, the corresponding latent vector representing the channel effects for a speech recording. Since $\mathbf{D}\mathbf{z}$ is trained using speaker-phrase sufficient statistics, $\mathbf{z}$ is said to capture the joint speaker-content information effectively thereby rendering the model more suitable for text-dependent speaker verification. As channel effects are modelled by $\mathbf{U}\mathbf{x}$, there is no need for a back-end model like PLDA in this case. In this paper, we use maximum-likelihood (ML) estimation algorithm [12] to obtain $\mathbf{D}$ and $\mathbf{U}$. We use the Gauss-Seidel approach [13, 14], maximizing the likelihood to obtain estimates of $\mathbf{z}$ and $\mathbf{x}$.

3. POSTERIORS FOR SPEAKER VERIFICATION

The posterior probabilities computed while estimating an i-vector or JFA factors assume feature vectors to be generated by a GMM. In the past, several studies have suggested that integrating linguistic information into speaker recognition systems can be useful [7, 15, 16, 17, 18]. In HMM/DNN automatic speech recognition [7], state posterior probabilities are obtained after ASR decoding. These are used to compute zeroth and first order statistics using the actual feature vectors of an utterance. This approach obtained significant improvements over a baseline i-vector-PLDA system. This suggests that i-vectors benefit from the acoustic space being partitioned by well-defined linguistic units. Clearly, this is difficult to achieve using unsupervised training, as used for GMM-UBM estimation.

After the successful integration of DNN posteriors into an i-vector PLDA text-independent system, we explored its application to text-dependent systems. Indeed, the very same approach can be readily applied to JFA systems as well.

3.1. HMM/DNN ASR system

In Automatic Speech Recognition (ASR), the acoustic models are context-dependent tied states [19], obtained using a decision tree based on contextual and data-driven criteria. A HMM/GMM system typically obtains the optimal state alignment for the training data, used to extract state labels for DNN training. The DNN using a final softmax layer aims at estimating the posterior probabilities of such tied states from a splice of input features. Given the large number of DNN outputs, in the thousands, the estimated posterior vectors tend to be sparse. A major drawback of training such a DNN is the need for a large amount of transcribed data. On the other side, posteriors for linguistic units are obtained.

For text-dependent speaker recognition, we believe state posteriors are particularly useful to capture the content variability. These are estimated using a supervised and discriminative procedure, rendering them more reliable than GMM posteriors, obtained fully in an unsupervised way.

In this paper, we use the state posterior probabilities from a DNN to estimate the zeroth and first order statistics for i-vector and JFA systems. In particular, the Baum-Welch statistics required to iteratively compute $\mathbf{z}$ and $\mathbf{x}$ use DNN posteriors. We also test the effectiveness of using these posteriors in a template matching system that uses Dynamic Time Warping (DTW), thereby modeling the sequence information. The details of the system architectures are provided next.

3.2. I-vector system from DNN posteriors

Using DNN posteriors in the i-vector system involves discarding the GMM-UBM entirely. However, the bias term ($\boldsymbol{\mu}$) in Equation 1 still needs to be estimated. This can be easily achieved by combining the posteriors and the corresponding feature vectors as follows: first, the components of the GMM-UBM are replaced by the states of the DNN. The mean ($\boldsymbol{\mu}_c$) and the covariance matrix ($\boldsymbol{\Sigma}_c$) of state c of the acoustic model are obtained from a development dataset. These parameters are obtained using the update equations for Expectation-Maximization of GMM from the raw features. This set of means and covariances serve as normalization factors in computing the zeroth and the first order statistics for i-vector extraction. The rest of the hyperparameter estimation process remains the same as in the conventional method, except that the posteriors are always computed using the DNN.

In this paper, we adapted this technique to the JFA (in Equation 2) as well. Once again, instead of using the traditional UBM, we use the mean and covariance parameters estimated from the posteriors from the DNN-ASR system. As the z-vector is shown to model speaker-phrase information, we expect the model to fully exploit the linguistic information supplied by the ASR system.

3.3. Template matching with posteriors

In the case of the DNN-based i-vector system, the sequence information provided by the DNN is not modeled. As a result, the text constraints imposed on the speaker are not fully exploited. In this work, we use the Dynamic Time Warping (DTW) algorithm as a scoring method, i.e. computing distances between target and test speaker utterances.

The DTW algorithm takes two sequences as input and matches their content by finding the path with the smallest alignment between them. For DTW to be used for text-dependent speaker recognition, it is sufficient that the algorithm is able to detect (i) different content being uttered by the same speakers (ii) different speakers speaking the same content. Most importantly, a template matching algorithm provides good benchmarks to ideal performance of the system when some information about the text-constraint can be assumed. For instance, in the case of speaker verification where it can be assumed that the speaker is the same but the content need not be, which is essentially speaker-dependent content matching, DTW has to just match the content. In such cases, sequence matching methods such as the DTW can be expected to perform the best.

In this paper, we use the posterior sequences obtained from the DNN for template matching. We hypothesize that such a system performs best for tasks on which content mismatch is treated as an impostor for speaker verification.

DTW-based template matching with DNN is performed as follows. Assume that an utterance contains M frames of speech. The DNN outputs a posterior vector for a multi-frame around each time instant. Each element in the posterior vector is the posterior probability of the state given the observations. The sequence of speech frames are represented by $\mathbf{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_\mathrm{M}\}$, where $\mathbf{o}_i$ is the i^{th} speech frame. The corresponding sequence of posterior vectors is

$\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_M\}$. In the literature, $\mathbf{P}$ represented as a matrix with $\mathbf{p}_m$ as columns is referred to as posteriograms. To compare two such sequences of posteriors the DTW algorithm minimizes the overall Kullback-Leibler (KL) divergence cost, taken as the distance between two sequences. Such an approach using only DNN posteriors can be expected to perform well when the content information is the same from train to test.

4. EXPERIMENTAL SETUP

In this section, we describe the experimental setup for the baseline and the proposed systems, and the system configuration of the i-vector-PLDA system, JFA system and the HMM-DNN ASR system.

4.1. Evaluation data

The experiments are conducted on the RSR2015 database, which is designed specifically for text dependent speaker verification task. The experiments are performed only on the female speaker set of the Part1 of the database. The Part1 consists of 30 fixed pass phrases and the duration of the utterances varies from 3s to 4s. The enrollment condition consists of 49 female speakers and 3 samples for each of the 30 phrases. The speaker verification systems are evaluated in three conditions. In condition 1, each trial is associated in determining if the phrases are the same or different. In condition 2, the system is required to differentiate speakers saying the same content. In condition 3, both the speaker and the phrase can be different. There are a total of 47 target speakers over the 30 different phrases. The dataset also contains a development set with 49 female speakers that can be used to train or adapt hyperparameters of the systems. All speech files are downsampled to 8kHz for compatibility with other datasets used for system development.

4.2. i-vector-PLDA and JFA system configurations

MFCC features with 20 dimensions are extracted from the speech signal along with delta and acceleration parameters. Short time gaussianization is applied to the features using a 3 sec sliding window [20]. A subset of Fisher database (approximately 120 hours) of female speech utterance is used to train the parameters of a 1024 mixture UBM and i-vector system ($\mathbf{T}$) of 400 dimensions. To train the PLDA model, the development data for Part1 of the RSR2015 database is used.

To train the JFA system only the development data from the RSR2015 dataset is used as it is necessary to have multiple sessions of speaker-phrase combinations. The UBM is obtained by adpating the UBM trained on Fisher dataset [21]. The eigenchannel matrix was trained with rank 50. The trails are evaluated by a simple cosine distance scoring. Unlike in [11], in which the JFA systems are evaluated in only the Condition 2, we test our systems on all conditions.

4.3. DNN system

The HMM/DNN system is bootstrapped with alignments from a HMM/GMM based ASR system trained on context dependent phoneme units. The ASR systems have 1909 tied states. The DNN is configured with 4 hidden layers trained on MFCCs with a 11-frame context. The entire training is done on the subset of the Fisher corpus as mentioned earlier. The Word Error Rate (WER) of the ASR system is 24.7% when tested on a separate subset of the Fisher coprus with 720 utterances. Unlike the state-of-the-art approaches, the speaker independent DNN is trained. That is, techniques such as Feature space Maximum Likelihood Linear Regression (fMLLR) are not used because of limited adaptation data for speakers in the evaluation dataset.

The posteriors for the proposed systems are obtained at the output of a forward-pass on the DNN. For the i-vector and JFA systems, each posterior vector is processed to obtain the top 10 scoring states.

Table 1: Performance of all the systems on the RSR2015 database in terms of EER(%). The overall EER refers to the system performance across all the 3 conditions.

Systems/Conditions	#1	#2	#3	Overall EER
Baseline systems				
i-vector PLDA	1.2	3.0	0.3	0.9
JFA	1.6	2.3	0.5	0.8
GMM-posteriors with DTW	0.5	7.2	0.2	1.7
Proposed systems				
DNN-ivector PLDA	**0.8**	**2.5**	**0.2**	**0.7**
DNN-JFA	**0.1**	**1.0**	**0.02**	**0.24**
DNN-posteriors with DTW	**0.1**	8.4	0.1	2.0

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

Six systems are evaluated and compared on the conditions mentioned in Section 4.1:

- **i-vector PLDA**: the conventional i-vector-PLDA system for speaker recognition. This system is used as our baseline.
- **DNN-ivector PLDA**: the i-vector PLDA system that uses posteriors obtained from HMM/DNN ASR.
- **GMM-posteriors with DTW**: Posteriograms obtained from the UBM-GMM are compared using the DTW algorithm in this system.
- **DNN-posteriors with DTW**: This system uses Posteriograms obtained from the HMM/DNN ASR and compares two sequences usin the DTW algorithm.
- **JFA**: This system is an alternative baseline to the i-vector PLDA. It models speakers as given by Equation 2.
- **DNN-posteriors with JFA**: This system uses posteriors from the ASR system instead of the conventional UBM-GMM models.

Table 1 compares the performances of all above-mentioned systems across all 3 conditions in terms of Equal Error Rate (EER). The performances on the combined condition are also presented as the "Overall EER". EER for the model-based baselines, namely the i-vector PLDA and the JFA system, are comparable (better in most conditions) to those found in the literature. The baseline for the DTW based template matching system uses the posteriors obtained from the GMM-UBM system, both for i-vector PLDA and JFA. In conditions 1 and 3, the GMM-DTW system is better compared to other two baselines as it explicitly matches the content. Its performance serve as a good benchmark for those conditions. In all three conditions, the JFA system outperforms the i-vector PLDA system. This validates the assumption that low rank eigenvoice modelling in JFA ($\mathbf{V}$ matrix in [22]), which is similar to Equation 1, does not

capture the content information as well as the model in Equation 2. Among the baseline systems, the JFA system provides the best overall performance. Therefore, in the following text unless mentioned we compare the overall EER only with the JFA baseline.

Incorporating DNN posteriors from HMM/DNN into the i-vector-PLDA and JFA systems leads to consistent improvements. For the DNN-i-vector-PLDA, the overall EER improves by 12% relative (0.7% vs. 0.8%). Improvements observed with the DNN-JFA system are far superior to all other gains achieved. The DNN-JFA system outperforms the baseline by 70% relative EER (0.24% vs. 0.8%) and is the best system among the proposed in all conditions. Thus, the combination of using a speaker-content model (Equation 2) along with leveraging linguistic information system is highly essential for accurate text-dependent speaker verification.

On conditions 1 and 3, the DTW based approaches are amongst the best ranked systems. For these conditions, train and test contents are different, the DTW algorithm is sufficient to discriminate impostor from target speakers. However, this can not be applied to all conditions, condition 2 in particular. The use of DNN-based posteriors once again helps improve the system to detect content mismatch with relative improvements of 80% (0.1% vs. 0.5%) and 50% (0.1% vs. 0.2%) EER on conditions 1 and 3, respectively.

In general, the use of a DNN system for extracting posteriors is observed to be useful. The gains obtained by incorporating the DNN in the two different speaker recognition frameworks supports the hypothesis that exploiting content information is essential to achieve high recognition accuracies.

6. CONCLUSIONS

The problem of text-dependent speaker verification was addressed. Three different systems were considered, namely the i-vector PLDA system, the JFA system and a simple DTW-based template matching system. In general, the JFA system performed better with the best overall EER of 0.8% among the three systems. As the model based approaches do not use the sequence information explicitly, the DTW-based approach performed better in conditions that required content mismatch detection. A DNN/HMM based ASR system was incorporated into the baseline systems to make better use of the content information. Significant performance gains are obtained in the model based approaches with the best EER of 0.24% with the DNN-JFA system. Thus, utilising sequence information obtained from ASR systems can be beneficial for text-dependent speaker verification.

7. ACKNOWLEDGEMENTS

This work was supported by the EU FP7 project Speaker Identification Integrated Project (SIIP).

8. REFERENCES

[1] Anthony Larcher, Kong-Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[2] A. Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Modelling the alternative hypothesis for text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 734–738.

[3] N. Dehak, P. 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, May 2011.

[4] P. Kenny, T. Stafylakis, P. Ouellet, and M.J. Alam, "Jfa-based front ends for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 1705–1709.

[5] Joseph P Campbell Jr, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[6] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.

[7] Yun Lei, N. Scheffer, L. Ferrer, and M. McLaren, "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*, May 2014, pp. 1695–1699.

[8] Daniel Garcia Romero and Carol Y. Espy Wilson, "Analysis of ivector length normalization in speaker recognition systems," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27 to 31, 2011*, 2011, pp. 249–252.

[9] O Glembek et al., "Simplification and optimization of i-vector extraction," 2011, pp. 4516–4519, In Proc. of ICASSP.

[10] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, and Mohammad Jahangir Alam, "Jfa-based front ends for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1705–1709.

[11] Patrick Kenny, Themos Stafylakis, J Alam, Pierre Ouellet, and Marcel Kockmann, "Joint factor analysis for text-dependent speaker verification," *Odyssey*, 2014.

[12] Patrick Kenny, G Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 335–354, May 2005.

[13] Robert J Vogt, Brendan J Baker, and Sridha Sridharan, "Modelling session variability in text independent speaker verification," 2005.

[14] Robbie Vogt and Sridha Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[15] Petr Motlicek et al., "Employment of subspace gaussian mixture models in speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4445–4449.

[16] Alex Park and Timothy J Hazen, "Asr dependent techniques for speaker identification.," in *INTERSPEECH*, 2002.

[17] Douglas E Sturim, Douglas A Reynolds, Robert B Dunn, and Thomas F Quatieri, "Speaker verification using text-constrained gaussian mixture models," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–677.

[18] Brendan J Baker, Robert J Vogt, and Sridha Sridharan, "Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification," 2005.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer, "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.

[20] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," 2001, pp. 213–218, In Proc. of Speaker Odyssey.

[21] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Diigital Signal Processing*, vol. 10, pp. 19–41, 2000.

[22] P. Kenny, G .Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, 2005.