JFA MODELING WITH LEFT-TO-RIGHT STRUCTURE AND A NEW BACKEND FOR TEXT-DEPENDENT SPEAKER RECOGNITION

Patrick Kenny[1], Themos Stafylakis[1], Jahangir Alam[1] and Marcel Kockmann[2]

[1]Centre de Recherche Informatique de Montreal (CRIM), Quebec, Canada
[2]VoiceTrust, Ontario, Canada

patrick.kenny@crim.ca

ABSTRACT

This paper introduces a new formulation of Joint Factor Analysis (JFA) for text-dependent speaker recognition based on left-to-right modeling with tied mixture HMMs. It accommodates many different ways of extracting multiple features to characterize speakers (features may or may not be HMM state-dependent, they may be modeled with subspace or factorial priors and these priors maybe imputed from text-dependent or text-independent background data). We feed these features to a new, trainable classifier for text-dependent speaker recognition in a manner which is broadly analogous to the i-vector/PLDA cascade in text-independent speaker recognition. We have evaluated this approach on a challenging proprietary dataset consisting of telephone recordings of short English and Urdu pass-phrases collected in Pakistan. By fusing results obtained with multiple front ends, equal error rate of around 2% are achievable.

Index Terms— Joint Factor Analysis, text-dependent speaker recognition

1. INTRODUCTION

We have developed a general version of Joint Factor Analysis (JFA) which we refer to as multi-tier JFA and which is designed to tackle several problems that arise in text-dependent speaker recognition but not in text-independent speaker recognition. This model opens up a large number of avenues for experimentation in modeling speaker effects, channel effects and for transfer learning. In this paper we will present an overview of multi-tier JFA and a new backend that we have developed to go with it. We report the results of a preliminary exploration of the possibilities afforded by multi-tier JFA for modeling speaker effects on a telephone-based, text-dependent speaker recognition task using proprietary data collected in Pakistan.

The most obvious difference between the text-independent and text-dependent tasks is that the text-dependent problem has a left-to-right structure which is more naturally modeled by HMMs than by GMMs. This suggests that in modifying JFA to handle text-dependent speaker recognition, it is necessary to introduce *local* hidden variables which vary from one HMM state to another in order to model speaker effects. Local hidden variables are difficult to handle because of data fragmentation (utterances are already very short in text-dependent speaker recognition) so *global* hidden variables common to all HMM states need to be accommodated as well. The same problem arises in conventional HMM modeling for text-dependent speaker recognition. The *HiLam* architecture in [9] implicitly uses hidden variables of both types in creating a speaker-dependent HMM for a target speaker at enrollment time.

Another major difference between the two tasks is that text-dependent data for background modeling is so hard to come by that subspace methods for modeling speaker effects are not nearly as effective in text-dependent speaker recognition as in the text-independent situation.[1] We will present some good results on subspace modeling of speaker effects in this paper but it is generally agreed that it is necessary in practice to use supervector-sized features in conjunction with subspace features to characterize speaker effects in text-dependent speaker recognition. For example, relevance MAP has to be combined with speaker factors in [4] and features extracted from the orthogonal complement of the total variability space have to be combined with i-vectors in [2]. We referred to these two types of hidden variable as $\boldsymbol{z}$-vectors and $\boldsymbol{y}$-vectors in our earlier work [3] (this notation is borrowed from the initial formulation of JFA [15]). In redesigning JFA, we need to allow for the possibility of using both $\boldsymbol{y}$ and $\boldsymbol{z}$ vectors for modeling local and global speaker effects in text-dependent speaker recognition.

A third consideration is that because of the dearth of text-dependent background data, text-independent resources such as Switchboard and Mixer need to be brought to bear on text-dependent speaker recognition problem. This type of transfer learning is beyond the scope of the current paper but we men-

[1]Using subspace methods to model channel effects in text-dependent speaker recognition is easier [3]. This is to be expected since channel effects are presumably much the same in both cases. On the other hand talking of speaker effects in the context of text-dependent speaker recognition is a bit misleading as the classes to be recognized are speaker-phrase combinations rather than speakers as such.

tion it here as it provides further motivation for developing a new, flexible JFA framework. Although we need to explore other methods such as those proposed in [2], we found in [3] that a simple type of Universal Background Model (UBM) adaptation to individual pass-phrases enables JFA modeling to operate in a pass-phrase independent way. (This idea extends straightforwardly to the case where pass-phrases are modeled by HMMs provided that a tied-mixture structure is used.)

However, our experiments in [3] showed that traditional speaker factors ($\boldsymbol{y}$-vectors in our current terminology) extracted from a JFA model trained on text-independent data were only moderately successful. This is not surprising since the success of text-dependent speaker recognition with very short utterances depends critically on modeling speakers' pronunciation of individual words whereas traditional speaker factors characterize the configuration space of speakers' vocal tracts. But it does suggest that an extended version of JFA ought to be able to handle *both* types of speaker modeling.

Although we will not pursue the issue in this paper, it seems obvious that modeling channel or session effects in text-dependent speaker recognition also stands to benefit from a more flexible JFA framework. These considerations led us to formulate the "multi-tier" version of JFA which we outline in Section 2. We explain how this model can be used to extract features to characterize speakers from enrollment and test utterances in a variety of ways. In Section 3, we describe the new Joint Density Backend (JDB) that we are using to perform speaker recognition with these features. Finally, in Section 4, the results are presented and compared to a baseline GMM-UBM system, yielding about 30% relative improvement in DCF.

2. MULTI-TIER JFA AS FEATURE EXTRACTOR

2.1. Baum-Welch statistics

Our starting point is a conventional universal background model which may be trained on text-independent data such as the Mixer corpora or on heterogenous text-dependent data. For each passphrase we construct a Gaussian codebook by performing several iterations of relevance MAP as in [3] and we use this to build a *speaker-independent* left-to-right tied mixture model. (For the experiments reported here we used codebooks of size 128 and 5 state HMMs.) The fact that these tied mixture models are ultimately derived from a common codebook enables us to build JFA models common to all pass-phrases.

We use the speaker-independent HMMs to segment utterances and to collect Baum-Welch statistics from each segment (i.e. HMM-state) in each utterance. (Our experience has been that using speaker-adapted HMMs for this purpose leads to slight degradations in performance.)

2.2. Hidden variables

The role of JFA is to define a joint probability distribution on collections of utterances of a given phrase by a given speaker. The hidden variables which are used to model speaker effects serve as features for speaker recognition. (A feature extracted from enrollment utterances is compared with the corresponding feature extracted from a test utterance using the Joint Density Backend described in Section 3.)

In an N-tier JFA model, each segment has an N-tuple of hidden variables associated with it. Depending on how the hidden variables are tied across recordings and segments, they may serve to model speaker effects or channel effects and these effects may be local or global.

To take a concrete example, JFA as originally formulated had three tiers, of which two modeled speaker effects and one channel effects. In [15] the expression

$$\boldsymbol{s}_r = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_r + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z} \tag{1}$$

describes the supervectors corresponding to a collection of recordings by a given speaker indexed by r. The hidden variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are tied across recordings and serve to model speaker effects. The hidden variables $\boldsymbol{x}_r$ vary from one recording to another and serve to model channel effects. There is no notion of "segment" in this situation as the original formulation of JFA was designed for text-independent speaker recognition. Note that each hidden variable is lifted to a supervector in (1) ($\boldsymbol{x}_r$ lifts to $\boldsymbol{U}\boldsymbol{x}_r$ etc.)

In the case of an N-tier model, a supervector is associated with each segment by lifting the corresponding hidden variables. Comparing these supervectors with the Baum-Welch statistics for all segments in all recordings enables us to infer the values of the hidden variables for the given collection of utterances by the given speaker and so to extract the features which serve to characterize the speaker. Vogt's Gauss-Seidel algorithm [10] is the prototype for this type of calculation and it is an instance of the variational Bayes algorithm [7]. We will present the variational Bayes posterior calculations and EM algorithms for the N-tier JFA model in detail elsewhere.

In this paper we will focus on using the N-tier JFA formalism for modeling local and global speaker effects using both $\boldsymbol{y}$ and $\boldsymbol{z}$ vectors. (Thus we will not explore the possibilities for modeling channel effects or transfer learning.) Tying hidden variables across recordings ensures that they serve to model speaker effects. If such hidden variables are *untied* across segments then they serve to model local speaker effects; on the other hand, tying across segments ensures that they model global speaker effects. In each case, speaker effects could be modeled with $\boldsymbol{y}$-vectors, $\boldsymbol{z}$-vectors or both. So there is a wide range of possibilities to explore even though we are focussing solely on modeling speaker effect.

3. THE JOINT DENSITY BACKEND

We are using hidden variables in multi-tier JFA that model speaker effects as features for speaker recognition. These hidden variables are tied across all recordings of a speaker so they always have the property that their number is independent of the number of recordings available to enroll the speaker. Thus speaker recognition can be carried out by comparing the hidden variables extracted from one or more enrollment recordings with the hidden variables extracted from a test recording. Cosine distance is perhaps the simplest possibility but we have found a probabilistic approach to be more effective.

3.1. Likelihood ratio

Given a pair of features X_e and X_t extracted from a given target speaker's enrollment data and a given test utterance we can form a likelihood ratio for speaker verification of the form

$$\frac{P_T(X_e, X_t)}{P_N(X_e, X_t)}$$

where P_T refers to the joint distribution of feature pairs occurring in target trials and P_N to the joint distribution in non-target trials. We can assume that $P_N(X_e, X_t)$ factorizes as $P_T(X_e)P_T(X_t)$ and concentrate on modeling the numerator. If the features are of low dimension (the case of $\boldsymbol{y}$ vectors), then we can model the numerator as a multivariate Gaussian and train it on a large number of target trials (which will generally involve heterogeneous pass phrases). If the features are supervector-sized we apply the same idea for each the mixture component seperately and sum-up the LLRs.

This idea is borrowed from [13]. It is actually more natural in the case of the features we use since we do not have to resort to i-vector averaging in order to equalize the number of features on the enrollment side and test side. As in the text-independent case, we observed that length normalization is important but, contrary to the text-independent case, score normalization is also important (especially in the case of supervector-sized features). For this purpose we use the same domain-dependent background set (comprising data collected from 100 speakers) that we used to train passphrase dependent codebooks for the Pakistan test set.

3.2. Gender averaging

We have found that is helpful *not* to use ground truth gender information concerning target and test speakers, but to follow the method proposed in [14] instead. Given enrollment and test features X_e and X_t, using m to refer to male and f to female, we first calculate posteriors $\pi(m|X_e), \pi(m|X_t), \pi(f|X_e)$ and $\pi(f|X_t)$ using a GMM-based gender classifier. We then calculate the score of the trial in two ways, once using a tied mixture model trained only on male speakers and similarly for the female case, obtaining two normalized scores S_m and S_f. We obtain a final score by weighting these as follows:

$$\pi(m|X_e)\pi(m|X_t)S_m + \pi(f|X_e)\pi(f|X_t)S_f \qquad (2)$$

Note that the weights of each score do not sum-up to one. In fact, in case where the trial is cross-gender, both weights should normally tend to zero, i.e. to the mean of the non-target scores, since S_m and S_f are score-normalized.

Dataset	G	EER (%)	DCF_{08}	DCF_{10}
RSR	m	7.6×10^{-3}	0.38×10^{-3}	0.72×10^{-3}
”	f	18.6×10^{-3}	0.86×10^{-3}	1.15×10^{-3}
Pakistan	m	2.45	0.091	0.287
”	f	3.77	0.153	0.370

Table 1. Comparison between RSR2015 (Part 1) and Pakistan dataset (english phrase). A fusion of 11 GMM-UBM system with SNorm is used, where the fusion weights have been trained on the test set of each dataset independently of each other.

4. EXPERIMENTS

4.1. Experimental set-up

The algorithms are scored against a proprietary dataset collected from landline and mobile telephony of Pakistan that consists of two different phrases (the first in English and the second one in Urdu). A summary of the dataset is given in Table 2. Cross-gender trials are included and the gender averaging technique is applied. We enroll speakers using three repetitions of the same phrase, coming from the same session, while enrollment and test utterances are from different sessions.

To provide an estimate of the difficulty of the dataset, we compare results of a baseline GMM-UBM system with fusion of 11 multiple front-ends and VADs. In Table 1 we notice that while on RSR2015 the system attains extremely low EER and DCF, on Pakistan data the EER are over the range of 2-4%. This results demonstrates the level of difficulty of the Pakistan dataset and how channelling would be to train the whole system using mostly out-of-domain data.

We have trained UBM, JFA and JDB using 3 text-dependent datasets, namely RSR2015, CSLU and a proprietary dataset collected at Concordia University. To train JFA, we are defining as class all utterances of the same speaker-phrase combination, independently of the session, so that it captures the channel variability. On the other hand, the joint-density model uses in the enrollment side a single $\boldsymbol{z}$- or $\boldsymbol{y}$-vector of the same speaker-phrase and session combination. 60-dimensional PLP with mean and variance normalization (MVN) are used as front-end features. In the case of fusion of

Phrase	#male	#female	#tar trials	#non trials
English	216	78	752	217868
Urdu	223	77	921	271362

Table 2. Statistics of the database (evaluation set). Number of speakers per gender and number of target and non target trials. Cross-gender trials are included.

Phrase	#States	EER	DCF_{08}	DCF_{10}
English	1	**1.82%**	**0.067**	0.232
”	5	2.07%	0.070	**0.186**
”	1&5 (single)	1.86%	0.066	**0.194**
”	1&5 (two)	**1.80%**	**0.065**	0.200

Table 3. Single vs. 5-state HMM vs. their fusion (1&5), on the English phrase. For fusion, we either use a single JFA system or two systems trained independently.

4 different systems, LFCC-MVN, MFCC-MVN and MFCC with short-term gaussianization are used as well. The systems are trained and evaluated independently of each other, apart from the voice detector and the segmentation of utterances into states, which are shared between all systems.

4.2. Experimental results

In the first set of experiments, we are training three different JFA models. Two with a single speaker tier (1-state and 5-state) and the third model with two tiers of the same type. The results for the english phrase are given in Table 3 and show that the HMM structure is useful for the law-false alarm area. Moreover, the last two lines show that the results obtained by fusing to JFA systems separately are very close to those obtained by a single JFA model with two tiers. Note that we are fusing the two systems using weights equal to 0.3 for the 5-state model and 0.7 for the single state.

In the next set of experiments we are fixing the model to be a single JFA with two speaker tiers and weight as before. We report results where cosine distance is used for back-end rather that the JDB, showing that JDB seems to have better performance. Moreover, a second JFA model is trained, where a 150-dimensional speaker subspace tier (i.e. $\boldsymbol{y}$-vectors) is deployed to model state-dependent hidden variables, yet with $\boldsymbol{z}$-vectors for modelling the whole utterance. Its performance is comparable to the first system. We note though that when both tiers are using $\boldsymbol{y}$-vectors, the degradation is severe. Finally, results using fusion of 4 different front-ends are given. In this case, a logistic regression model is trained for each gender separately using the Bosaris toolkit, with DCF_{08} as optimization criterion.

In the last set of results given in Table 5, the proposed method is compared to a baseline GMM-UBM system with SNorm. We have also excluded cross gender trials and evaluated the system with and without gender averaging (GA). The gain in performance over the baseline is significant, especially in the DCF metrics. Moreover, the use of GA is very beneficial for female speakers (due to the small number of female score normalization utterances) with the cost of a slight degradation in male speakers.

Phrase	System	EER	DCF_{08}	DCF_{10}
English	CosDist-$\boldsymbol{z}$-plp	**1.72%**	0.071	0.212
”	JDB-$\boldsymbol{y}$-plp	1.95%	0.075	**0.194**
”	JDB-$\boldsymbol{z}$-plp	1.86%	**0.066**	**0.194**
”	JDB-$\boldsymbol{z}$-fusion	1.94%	0.061	0.162
Urdu	CosDist-$\boldsymbol{z}$-plp	2.74%	0.109	**0.269**
”	JDB-$\boldsymbol{y}$-plp	**2.70%**	**0.103**	0.275
”	JDB-$\boldsymbol{z}$-plp	2.71%	**0.103**	0.273
”	JDB-$\boldsymbol{z}$-fusion	2.25%	0.091	0.239

Table 4. Comparison between Cosine Distance (CosDist) and Joint Density Model with Length Normalization, using either PLP or fusion of 4 front-ends and $\boldsymbol{z}$- or $\boldsymbol{y}$-vectors. The HMM is 5-state, while SNorm and gender averaging are applied.

G	System	GA	EER	DCF_{08}	DCF_{10}
m	baseline	n	2.29%	0.092	0.318
m	JDB-$\boldsymbol{z}$	n	**1.71%**	**0.065**	0.171
m	JDB-$\boldsymbol{z}$	y	1.97%	0.066	**0.163**
f	baseline	n	3.99%	0.198	0.451
f	JDB-$\boldsymbol{z}$	n	3.68%	0.138	**0.381**
f	JDB-$\boldsymbol{z}$	y	**2.82%**	**0.116**	0.418

Table 5. GMM-UBM (baseline) vs. proposed system on the english phrase and the effect of gender averaging (GA) per gender. SNorm is applied to all of the systems.

5. CONCLUSION

A text-dependent speaker recognition system was proposed, that combines JFA-features with a probabilistic classifier and HMMs for segmentation of the utterances. The core of the system is a JFA that is trained on out-of-domain data and requires a small amount of unlabelled in-domain data for UBM adaptation and score normalization. Moreover, the particular back-end fits well to the $\boldsymbol{z}$- and $\boldsymbol{y}$-features, while gender averaging can be deployed as an alternative to gender detectors.

The experiments on Pakistan telephony demonstrated the clear superiority over a GMM-UBM baseline, while the results can be improved by applying fusion of several front-ends. The performance of $\boldsymbol{y}$-vectors shows that speaker subspaces are applicable, at least when combined with $\boldsymbol{z}$-vectors. Finally, the system may also be explored using DNNs for extracting Baum-Welch statistics, a method that is needed to be examined in depth for the text-dependent case, [17] [18].

6. REFERENCES

[1] H. Aronowitz and O. Barkan, "On leveraging conversational data for building a text dependent speaker verification system," *Interspeech* 2013.

[2] H. Aronowitz and A. Rendel, "Domain Adaptation for Text-Dependent Speaker Recogntion," *Interspeech* 2014.

[3] P. Kenny, T. Stafylakis, J. Alam, P. Ouellet and M. Kockmann, "In-Domain versus Out-of-Domain Training for Text-Dependent JFA," *Interspeech* 2014.

[4] A. Miguel, J. Villalba, A. Ortega, E. Lleida and C. Vaquero, "Factor Analysis with Sampling Methods for Text Dependent Speaker Recognition," *Interspeech* 2014.

[5] T. Stafylakis, P. Kenny, *et al.*, "Text-dependent speaker recognition using PLDA with uncertainty propagation," *Interspeech* 2013.

[6] P. Kenny, T. Stafylakis, P. Ouellet, and M. J. Alam, "JFA-based front ends for speaker recognition," *ICASSP* 2014.

[7] P. Kenny, T. Stafylakis, M. J. Alam, P. Ouellet and M. Kockmann, "Joint Factor Analysis for Text-Dependent Speaker Verification," *Odyssey* 2014.

[8] A. Larcher, K.-A. Lee, B. Ma, and H. Li, "Phonetically constrained PLDA modeling for text-dependent speaker verification with multiple short utterances", *ICASSP* 2013.

[9] A. Larcher, K.-A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015 ", *Speech Communction*, (60), 2014.

[10] R. J. Vogt and S. Sridharan, "Explicit modeling of session variability for speaker verification," *Computer Speech and Language*, 2008.

[11] P. Kenny, G. Boulianne, *et al.* "Joint Factor Analysis versus eigenchannels in speaker recognition," *IEEE Trans. ASLP*, May 2007.

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. ASLP*, 2011.

[13] Sandro Cumani, and Pietro Laface, "Generative pairwise models for speaker recognition ," *Odyssey* 2014.

[14] M. Senoussaoui, P. Kenny, P. Dumouchel, N. Dehak, "New cosine similarity scorings to implement gender-independent speaker verification," *Interspeech* 2013.

[15] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms, Tech. Report CRIM-06/08-13," 2005. http://www.crim.ca/perso/patrick.kenny

[16] Cole, Ronald A., Mike Noel, and Victoria Noel. "The CSLU speaker recognition corpus," in *ICSLP*, Vol. 98. 1998.

[17] Y. Lei, N. Scheffer, L. Ferrer and M. McLaren, "A novel scheme for speaker recognition using a phonetically aware Deep Neural Network," *ICASSP* 2014.

[18] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet and J. Alam, "Deep Neural Networks for extracting Baum-Welch statistics for Speaker Recognition," *Odyssey* 2014.