In-Domain versus Out-of-Domain training for Text-Dependent JFA

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

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

patrick.kenny@crim.ca

Abstract

We propose a simple and effective strategy to cope with dataset shifts in text-dependent speaker recognition based on Joint Factor Analysis (JFA). We have previously shown how to compensate for lexical variation in text-dependent JFA by adapting the Universal Background Model (UBM) to individual passphrases. A similar type of adaptation can be used to port a JFA model trained on out-of-domain data to a given text-dependent task domain. On the RSR2015 test set we found that this type of adaptation gave essentially the same results as in-domain JFA training. To explore this idea more fully, we experimented with several types of JFA model on the CSLU speaker recognition dataset. Taking a suitably configured JFA model trained on NIST data and adapting it in the proposed way results in a 22% reduction in error rates compared with the GMM/UBM benchmark. Error rates are still much higher than those that can be achieved on the RSR2015 test set with the same strategy but cheating experiments suggest that if large amounts of in-domain training data are available, then JFA modelling is capable in principle of achieving very low error rates even on hard tasks such as CSLU.

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

10.21437/Interspeech.2014-330

1. Introduction

There has recently been a resurgence of interest in text-dependent speaker recognition for applications such as automated password reset and user authentication in financial service industries. The benefits compared to text-independent technology is that error rates similar to those obtained in recent NIST speaker recognition evaluations can be achieved with utterances of very short duration (typically 1-2 sec). The key element here is the matched phonetic content between enrollment and test utterances. By suppressing the phonetic variability between enrollment and test utterances, the system can focus on modelling speaker and channel variability, and achieve less than 1% EER in datasets like RSR2015, [5] [4].

An open question regarding text-dependent systems is how to leverage out-of-domain data (and especially NIST) to train models. Recall that the success of modern statistical methods of text-independent speaker recognition depends critically on the availability of the large development corpora that have been provided by NIST. Unfortunately, data for training text-dependent speaker recognition systems is still very limited and it is not clear to what extent the subspace methods that have proved to be so powerful in text-independent speaker recognition can be used in the text-dependent context.

Recently, we have developed a Joint-Factor Analysis (JFA, [10], [8]) approach to the text-dependent speaker recognition problem, [4]. On the RSR2015 Part I test set [6], this approach is capable of attaining error rates less than 1% when trained on in-domain data. In this paper, we show that the same system can be trained on NIST data -using a minimal amount of unlabelled in-domain data for adaptation and score normalization, with minimal degradation in performance compared to fully in-domain training. This indicates that NIST data can serve to model at least channel variability (but probably not speaker-phrase variability) in text-dependent speaker recognition. To the extent that this is true for a given application domain, a text-dependent speaker recognition system can be built without having to collect multiple recordings over different channels of speaker-phrase pairs.

It is generally agreed though that the channel variability of RSR2015 is not severe. This is partly due to the fact that all recordings of the same speaker were collected during the same day. Hence, we decided to examine whether the JFA adaptation approach could be used on a much more demanding text-dependent speaker recognition dataset, the CSLU speaker recognition corpus (specifically, on the 5-digit phrases), [11]. What is interesting with CSLU corpus is the fact that recordings of each speaker were collected over a two-year period. Thus, there is a severe aging effect that dominates "channel" variability, not appearing e.g. on NIST. There also appears to be substantial channel variability in the literal sense. Little appears to have been published on this test set and, contrary to the RSR2015 test set [6], the classical GMM/UBM approach seems to perform very poorly (see experiment). Our aim in this paper is to explore the use of factor analysis methods in text-dependent speaker recognition with particular emphasis on the problem of channel robustness. Although it is natural to use HMMs rather than GMMs to capture the left-to-right structure introduced by lexical constraints, working with HMMs would complicate channel modeling and experimentation without offering new insights into the problem. Thus we have chosen to work with GMMs rather than HMMs at this stage of our work. We show that although the results of JFA adaptation on this test set are far from being satisfactory, a relative improvement of 22% can be attained compared to a GMM/UBM system with t-norm, using the same NIST-trained JFA model as in our RSR2015 experiments.

The rest of the paper is organized as follows. In Section 2, an analysis of JFA is given, and the three different features we use are discussed. In Section 3, the experiments are presented, starting from those on RSR2015 and continuing with those on CSLU. Finally, the performance for each of the JFA-features is demonstrated, both for in-domain and out-of-domain JFA-training. For the in-domain training experiments in CSLU we used the enrollment data for all of the trials as JFA training material. We used a set-up similar to the 2012 NIST SRE but easier in that the test channels were exposed as well as the test speakers. The results are tantalising but, in the interests of

clarity, we emphasize that these experiments involve cheating in that they do not simulate the performance of an "application-ready" system.

2. JFA as a feature extractor

In this section we discuss the details regarding the use of JFA as a feature extractor, as well as phrase adaptation of the UBM and domain adaptation from NIST data to text-dependent datasets.

2.1. Analysis of JFA features

As in our previous work ([3], [4]), we use JFA to extract feature vectors for text-dependent speaker recognition. These features are fed into a simple back-end classifier (such as cosine distance with or without Within Class Covariance Normalization), [9]. Recall that the general JFA model assumes that, given a UBM with mean supervector $\boldsymbol{m}$ and multiple recordings of a speaker indexed by r, each recording can be modeled by a GMM whose (unobservable) mean supervector $\boldsymbol{S}^r$ has the form

$$\boldsymbol{S}^r = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}^r + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z} \tag{1}$$

where the hidden variables $\boldsymbol{x}^r, \boldsymbol{y}$ and $\boldsymbol{z}$ are assumed to have standard normal priors. The hidden variable $\boldsymbol{x}^r$ varies from one recording to another and is intended to model channel effects. In text-independent speaker recognition, the term $\boldsymbol{D}\boldsymbol{z}$ is usually dropped and speakers are characterized by the low-dimensional vector $\boldsymbol{y}$. For text-dependent speaker recognition, we drop the term $\boldsymbol{V}\boldsymbol{y}$ and we use the variables $\boldsymbol{z}$ to characterize speaker-phrase combinations. The prior on $\boldsymbol{z}$ is factorial in the sense that $P(\boldsymbol{z}) = \prod_c P(\boldsymbol{z}_c)$ where c ranges over mixture components and $\boldsymbol{z}_c$ is the part of $\boldsymbol{z}$ that corresponds to mixture component c. In the case of relevance MAP with relevance factor f, the corresponding submatrix $\boldsymbol{D}_c$ of $\boldsymbol{D}$ is defined by the condition that $f\boldsymbol{D}_c^*\boldsymbol{\Lambda}_c\boldsymbol{D}_c$ is the identity matrix where $\boldsymbol{\Lambda}_c$ is the precision matrix of the mixture component, [7].

The three different JFA-features (or simply JFA-vectors) are $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$. They are extracted by dropping certain terms from (1). To extract $\boldsymbol{x}$-features (i.e. the familiar i-vectors) we suppress both $\boldsymbol{V}\boldsymbol{y}$ and $\boldsymbol{D}\boldsymbol{z}$ and obtain a single low-dimensional vector for each recording. In this case, $\boldsymbol{U}$ is referred to as the total variability matrix, since it models both speaker and channel effects, [9].

On the other hand, $\boldsymbol{z}$-features are derived by dropping $\boldsymbol{V}\boldsymbol{y}$ only. Contrary to text-independent speaker recognition, the $\boldsymbol{z}$-features that are extracted from a collection of (enrollment or test) recordings characterize a speaker-phrase combination rather than a speaker as such. Channel effects are modelled and removed via the term $\boldsymbol{U}\boldsymbol{x}^r$. Similarly, to work with $\boldsymbol{y}$-features we need to suppress $\boldsymbol{D}\boldsymbol{z}$. Like $\boldsymbol{z}$, $\boldsymbol{y}$-features characterize a collection of recordings, meaning that during runtime, we extract a single $\boldsymbol{z}$ or $\boldsymbol{y}$-feature from the enrollment utterances (independently of their number) and one from the test utterance. A key difference between $\boldsymbol{y}$-features and $\boldsymbol{z}$-features is that $\boldsymbol{y}$-features can only be expected to work in situations where sufficient training data is available and subspace modeling methods can be applied. A speaker subspace trained on NIST data is unlikely to be useful in characterizing speaker-phrase combinations in a given text-dependent task domain. This is borne out in the experiments we report in this paper. For similar reasons, i-vectors have not fared very well in text-dependent speaker recognition, [2] [5] [6]. On the other hand, it is reasonable to assume that channel effects in text-dependent speaker recognition are the same as in text-independent speaker recognition and that these can be learned from NIST data. These considerations motivated our choice of $\boldsymbol{z}$-features for the work in [3] and [4]. In this paper we will report results obtained with all three types of feature sets on the CSLU corpus. Our results confirm that $\boldsymbol{z}$-features give the best results (except in the cheating experiments where JFA is trained on the enrollment data in our test set).

2.2. Phrase adaptation

In text-independent speaker recognition, utterances are usually long enough that the phonetic content is more or less averaged out so that phonetic variation is not a major source of nuisance variability. The situation is quite different in the text-dependent case, where utterances are typically of 1-2 sec duration. In [4] we showed how to compensate for this type of nuisance variability in JFA by using phrase-dependent background models (PBMs) to collect Baum-Welch statistics, instead of a single UBM as is conventionally done. Provided that the PBMs are all adapted from a UBM so that mixture components in the various models have a common interpretation (we used iterative relevance MAP), the JFA parameters which model channel variability can be shared across phrases. This serves to decouple channel variation from lexical variation and it results in a 50% reduction in error rates, [4]. Interestingly, passphrase background modelling does not help to improve the performance of a traditional GMM/UBM system (as we shall see).

2.3. Domain adaptation

Suppose now that we are given a JFA model and an associated UBM trained on out-of-domain data (NIST data in practice). All that is required to adapt the JFA model to a given text-dependent task domain is to produce PBMs by adapting from the UBM trained on NIST data, rather than from a UBM trained on within domain data.

Thus a limited amount of domain-dependent adaptation data is required for the approach we are proposing. Note that we do not require multiple recordings of speaker-phrase combinations to model channel effects in JFA. (Nor do we need such data for the backend classifier in the case of $\boldsymbol{y}$- and $\boldsymbol{z}$-vectors. On the other hand, i-vector classifiers can benefit from such data.)

It is interesting to note that our JFA-adaptation procedure works in the opposite direction from the one proposed in [1] (sect. 4). The starting point in that paper is a UBM trained on a text-dependent task domain (Wells Fargo). An i-vector extractor is trained from NIST data using Baum-Welch statistics extracted with the domain dependent UBM. This seems unnatural (although it works) and, if there are multiple pass-phrases (as in the case of the RSR2015 data) it would be unreasonable to proceed in this way for each PBM.

3. Experiments

3.1. Results on RSR2015

We begin the experiments by demonstrating results on RSR2015. For an analysis on the dataset, as well as the baseline performance we refer to [6]. Briefly, the dataset (Part 1) consists of 30 different phrases (taken from TIMIT), with durations ranging from 1 to 2 seconds. We use the background set (43 female and 50 male) for training and the evaluation set for testing (49 female and 57 male). The number of enrollment utterances is 3 and of the same handset, while the test utterance is of same phrase and from different handset. The results are

	Model	training	EER (%)	minNDCF
1	GMM/UBM	RSR	1.06	0.045
2	JFA	RSR	**0.61**	**0.027**
3	JFA	NIST	0.71	0.031

Table 1: Results of RSR2015 Part I female evaluation set.

	Model	training	EER (%)	minNDCF
1	GMM/UBM	RSR	0.60	0.034
2	JFA	RSR	**0.44**	**0.028**
3	JFA	NIST	0.58	0.031

Table 2: Results of RSR2015 Part I male evaluation set.

given in Table 1 and 2 in terms of Equal Error Rate and minimum normalized DCF (NIST 2008).

3.2. Results on CSLU

3.2.1. Experimental set-up

The sub-corpus of CSLU-SV-1.1 we used consists of only the six five-digit phrases. Each of these phrases is repeated 4 times per session, and there are 12 sessions per speaker in all. We reserved the fourth repetition of each phrase, for all sessions, for the test set. Due to the small size of the corpus, the first three repetitions were used for both enrollment and training JFA, or simply adapting the UBM, in the case of out-of-domain training. The overall number of speakers is 91, while the trial statistics are as given in Table 1. All trials are same-gender and same-phrase, while sessions with fewer than three repetitions were discarded.

3.2.2. Experimental Results

As our baseline, a standard GMM/UBM is deployed with t-norm. For enrolling a speaker-phrase model, a single MAP iteration is performed, while the log-likelihood ratio (LLR) is normalized with the number of frames of the test utterance. We have also experimented with using PBMs, so that we obtain a fair comparison between GMM/UBM and JFA-features. Interestingly, the performance was degraded, as Table 4 shows.

3.3. Results using JFA trained on NIST

We now focus on the three flavours of JFA-features, where the JFA model is trained on NIST data. CSLU is only used in order to adapt the UBM to PBMs and for score normalization. We applied 5 EM iterations with relevance factor equal to 2, while only the means are adapted. Note that in order to perform this adaptation, multiple recordings of the same speaker-phrase are not required. Thus, it can be considered as a realistic scenario for building a text-dependent speaker recognition system.

The results are given in Table 5, while the corresponding DET curves in Fig. 1 and 2 for female and male, respectively.

Gender	female	male
Speaker-phrase model	3304	3054
Test utterances	3303	3054
Target trials	35511	32480
Nontarget Trials	1780270	1518978

Table 3: CSLU trials in numbers.

	UBM training	PBM	EER (%)	minNDCF
1	CSLU	no	**6.83**	**0.253**
2	CSLU	yes	8.91	0.267
3	NIST	yes	8.48	0.266

Table 4: CSLU female, GMM/UBM with t-norm.

	JFA-feat.	PBM	Gender	EER (%)	minNDCF
1	$\boldsymbol{x}$	yes	female	7.87	0.311
2	$\boldsymbol{y}$	yes	female	6.61	0.263
3	$\boldsymbol{z}$	no	female	8.09	0.311
4	$\boldsymbol{z}$	yes	female	**5.76**	**0.228**
5	$\boldsymbol{x}$	yes	male	8.30	0.323
6	$\boldsymbol{y}$	yes	male	6.51	0.265
7	$\boldsymbol{z}$	no	male	6.22	0.254
8	$\boldsymbol{z}$	yes	male	**4.62**	**0.190**

Table 5: CSLU, JFA features trained on NIST, cosine distance with s-norm.

Clearly, the results show that $\boldsymbol{z}$ and $\boldsymbol{y}$ are superior to $\boldsymbol{x}$-features (i.e. i-vectors), when out-of-domain datasets are used for training. Moreover, the attempt to estimate a subspace for capturing speaker-phrase variability seems to be unsuccessful, since $\boldsymbol{z}$-vectors performed better that the $\boldsymbol{y}$-vectors. Finally, the benefits from adapting the UBM to the phrase are evident, yielding about 27% relative improvement in equal error rate (EER), when $\boldsymbol{z}$-vectors are deployed (compare lines 3 and 7 to 4 and 8 respectively).

3.4. In vs. out-of-domain training

We now show the results obtained when trained on CSLU versus on NIST data. We emphasize that the training set of CSLU coincides to the enrollment utterances, so that the model is vulnerable to overfitting the data. Moreover, all speakers and channels that appear on the test set are included in the enrollment set.

We start with the $\boldsymbol{x}$-features, i.e. the familiar i-vectors, extracted with PBMs. We do not apply PLDA in this paper but cosine-distance scoring followed by s-norm. Averaging of the enrollment i-vectors is performed using unnormalized i-vectors, followed by within-class covariance normalization (WCCN), cosine distance and s-norm. The results are demonstrated in Table 6. We note that while WCCN is helpful when trained on CSLU, it seems to be harmful when trained on NIST. This is an indication that channel modelling for CSLU based on NIST data is not feasible, when the i-vector approach is deployed.

The next set of experiments is performed with $\boldsymbol{y}$-vectors. Contrary to i-vectors, a single $\boldsymbol{y}$-vector is extracted from the

	training	WCCN	EER (%)	minNDCF
1	CSLU	no	3.98	0.159
2	CSLU	CSLU	**2.64**	**0.113**
3	NIST	no	7.87	0.311
4	NIST	CSLU	4.81	0.207
5	NIST	NIST	8.38	0.323

Table 6: CSLU female, JFA $\boldsymbol{x}$-features (i.e. i-vectors) with averaging, cosine distance and s-norm.

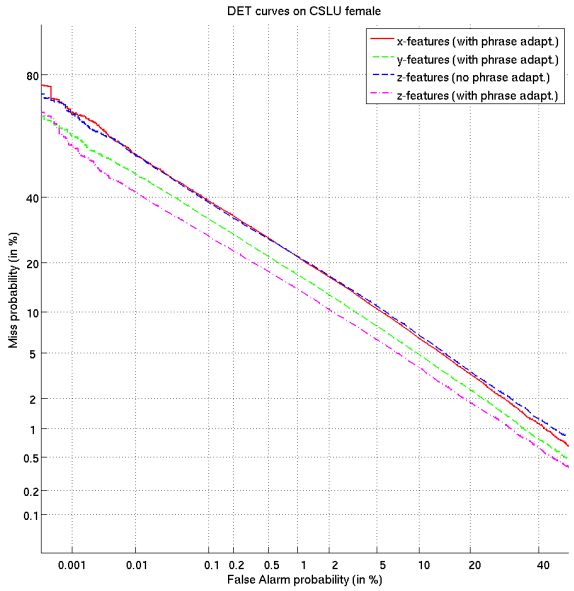

Figure 1: DET curves for the 3 flavours of JFA features on CSLU female, when JFA is trained on NIST.

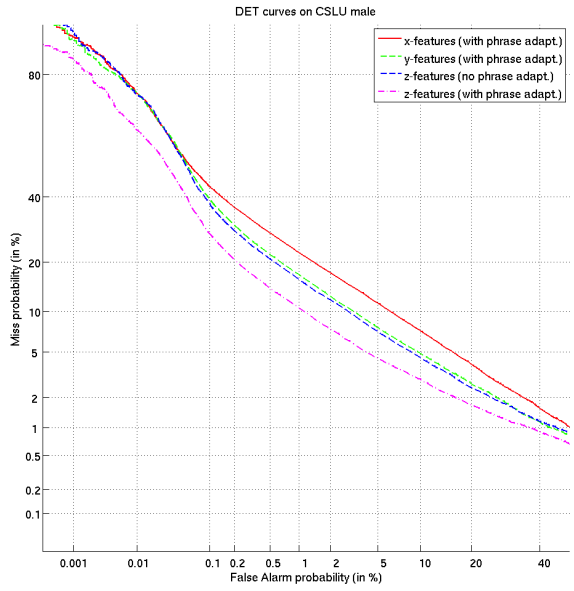

Figure 2: DET curves for the 3 flavours of JFA features on CSLU male, when JFA is trained on NIST.

	training	$\boldsymbol{y}$-dim	$\boldsymbol{x}$-dim	EER (%)	minNDCF
1	CSLU	100	50	0.84	0.032
2	CSLU	300	100	**0.35**	**0.012**
3	NIST	400	200	6.61	0.263

Table 7: CSLU female, JFA $\boldsymbol{y}$-features with s-norm.

	training	Gender	WCCN	EER (%)	minNDCF
1	CSLU	female	no	**3.03**	**0.129**
2	NIST	female	no	5.76	0.228
3	CSLU	male	no	**2.45**	**0.101**
4	NIST	male	no	4.62	0.190

Table 8: CSLU female & male, JFA $\boldsymbol{z}$-features with s-norm.

enrollment utterances. The results are given in Table 7. The dimensionality of the $\boldsymbol{y}$-vector is denoted by $\boldsymbol{y}$-dim, while $\boldsymbol{x}$-dim refers to the rank of the channel subspace. As the results show (line 1 and 2), training JFA on CSLU may lead to extremely low error rates, depending on the size of the speaker-phrase subspace. We should keep in mind though that all the variability of the dataset (excluding the test utterances) has been used to attain these results. Thus, overfitting is the major cause for this tremendous reduction in the error rates. Once we have trained the JFA model on NIST, the error rates were degraded by an order of magnitude, and became inferior to those attained with $\boldsymbol{z}$-vectors (Table 8, line 3). Hence, this result confirms the inadequacy of NIST data in estimating a speaker subspace that can be used as speaker-phrase subspace for CSLU.

Finally, we present the experiments with $\boldsymbol{z}$-vectors. The results are given on Table 8. In lines 1 and 3, UBM and JFA are trained on CSLU enrollment utterances, while in 2 and 4 on NIST. The only operation where CSLU utterances are used is for adapting the UBM to PBMs. It is interesting to note that although the error rates are nearly doubled when NIST is used for JFA training, the performance is still much better compared to our baseline GMM/UBM (22% relative improvement).

4. Conclusions

In this paper, we have described three JFA-based approaches to text-dependent speaker recognition and we have proposed a simple, effective strategy for adapting JFA models from one domain to another. Based on the encouraging results we obtained on the RSR2015 dataset, we chose to work on a more difficult dataset (CSLU) and evaluate the three JFA-based features. The improvement we attained over the baseline system was significant when $\boldsymbol{z}$-features were deployed (22% relative improvement). A key ingredient is the adaptation of the UBM to each phrase, that requires a minimal amount of in-domain unlabelled utterances. This single operation resulted in a 27% relative improvement over $\boldsymbol{z}$-vectors extracted with a JFA model trained on NIST data and no UBM adaptation. Finally, we reported results when JFA was trained on in-domain data. The superiority of $\boldsymbol{y}$-vectors over both $\boldsymbol{z}$ and $\boldsymbol{x}$-vectors was clear, attaining EER way below 1%. Of course, these results are not indicative of the performance of an application-ready system (since the enrollment utterances were included in the JFA training set). However they do suggest that if a system has been deployed for some time (so that large amounts of in-domain data can be collected), then subspace methods may prove to be just as effective in text-dependent as in text-independent speaker recognition.

5. References

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

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

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

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

[5] 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.

[6] A. Larcher, A.-K. Lee, B. Ma, H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015", *Speech Communication*, March 2013.

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

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

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

[10] 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

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