ACROSS-PHONE VARIABILITY AND DIAGONAL TERM IN JOINT FACTOR ANALYSIS FOR SPEAKER RECOGNITION

Sachin S. Kajarekar

SRI International, Menlo Park, CA, USA

ABSTRACT

We investigate usefulness of across-phone variability for speaker recognition in a joint factor analysis (JFA) framework. We estimate the variability as across-phone covariance within a conversation side averaged over all conversations. Note that it is a part of channel variability in the current JFA framework. We independently estimate feature subspaces representing across-phone, speaker and channel variability and perform speaker recognition experiments by either keeping them or removing them. The results show that the across-phone subspace is more correlated with the speaker subspace. We also perform speaker recognition experiments when combining the subspaces. Results show an improvement when phone and speaker subspaces are combined. This shows that across-phone variability is useful for speaker recognition. Further experiments show that the results are affected by a diagonal term from JFA. In particular, the improvement when combining the speaker and phone subspaces is reduced when the diagonal term is estimated from a universal background model (UBM). This implies that there is an interaction between the variability represented by the diagonal term and the across-phone variability. Overall, the work shows the importance of understanding the diagonal term (with speaker and channel subspaces) for incorporating additional variability into JFA beyond speaker and channel.

Index Terms**—** Speaker recognition, joint factor analysis, phonetic variability, language independent speech recognition

1. INTRODUCTION

Joint factor analysis (JFA) [1] is the most successful model of speaker and channel variability for speaker recognition. The initial model has been extended to include other factors, such as language and conversation segments [2].

Our previous work [3] focused on incorporating phonetic variability in a speaker recognition system. We showed significant improvements using polynomial cepstral features with the support vector machine (SVM) framework. The main idea was to train the channel or intersession variability matrix using all the data in the conversation and divide the compensated features into different phone categories. Independent SVM systems were trained for each class and were combined at the score level.

We performed similar experiments with JFA, cepstral features, and the Gaussian mixture model (GMM) framework. The results were consistent with those defined in the earlier paper [3], where global estimation of the speaker and channel factors gave the best performance (equal error rate, EER) over an independent estimate for each phone. However, a score-level combination of per-phone systems did not show any improvement.

Significant research was performed as part of a 2008 Johns Hopkins University summer workshop to incorporate phone-specific information in JFA for speaker recognition [4]. Results showed that the best performance was obtained by describing the speaker and channel subspace as stacking of phone-specific subspaces. In similar work [2], within-session variability was modeled in addition to conventional across-session variability in the JFA framework. The hypothesis was that removing within-session variability would improve the performance on short test utterances and was supported by the results.

One issue with the earlier approaches [3, 4] is that across-phone variability is ignored in factor analysis and is modeled at the score level. In this work, we focus on across-phone variability using factor analysis. The variability is estimated within each conversation side and averaged over all conversations. This is similar to earlier work [2] where fixed-length segments were used instead of phonetic classes.

There are many ways of incorporating across-phone variability into JFA. In this paper, we are exploring an approach to understanding the variabilities that are modeled in the conventional JFA framework. We want to know how similar a new factor is to existing factors and what would be the best way of incorporating the new variability into JFA.

2. EVALUATION FRAMEWORK

NIST 2008 Speaker Recognition Evaluation (SRE) data is used for the evaluation. We use short2-short3 condition 6, which has telephone conversation data from different languages. It has 1788 speaker models, 2569 tests, and 35869 trials. We further subset the results to English-only trials where both train data and test data are in English. These represent 17761 trials (also referred to as condition 7). All the results in this paper are reported in terms of equal error rate (%EER). Phone alignments were obtained from the Hungarian open phone loop recognizer from Brno University of Technology (BUT). The alignments were shared with the participants in the 2008 Johns Hopkins University summer workshop.

3. BASELINE JFA SYSTEM

The baseline system uses 13 Mel frequency cepstral coefficients [MFCCs] (C0-C12). They are processed with cepstral mean subtraction and have delta and double delta coefficients appended. A conversation side is segmented with SRI's hidden Markov model (HMM) based speech/silence segmenter. Speech segments are selected, and the 10% of frames with lowest energy are discarded. The remaining feature vectors are normalized by the mean and the variance of the individual features computed over the utterance.

A 1024-component GMM is trained on 2004 SRE data and is used as a universal background model (UBM). JFA is performed with the same data to estimate 300 speaker and channel factors. We use dot-product scoring [5] with ZTnorm score normalization where impostors are drawn from the 2004 and 2005 alternate microphone SREs.

In JFA [6], the assumption is that a given speaker (with a given channel) supervector m (with dimension NF = number of Gaussians(N)*number of features(F)) can be decomposed into a

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

sum of two supervector components: the speaker supervector Vy and the nuisance (or channel) supervector Ux and UBM mean m_0 ,

$$m = m_0 + Ux + Vy + Dz \cdot$$

The nuisance supervector distribution lies in a low-dimensional subspace of rank R_c, and is assumed to be distributed according to UU^T . Similarly, speaker supervector distribution lies in a low-dimensional subspace of rank R_s , and is assumed to be distributed according to VV^T. The matrix U (channel subspace) has a dimension of $NF * R_c$and the matrix V (speaker subspace) has a dimension of $F * R_s$. The subspaces U and V are estimated from a sufficiently large data set while the latent variables$x$ and y are estimated for each utterance.

D is interpreted in two ways. It can be coupled with *Vy* and the combined term can be referred to as a *speaker supervector*. Or *D* can be interpreted as an error term. It is the error in the approximation of *m* by m_0, *U,* and *V* that is expressed as a diagonal vector. It can be estimated in two ways and the choice can significantly affect the performance. In the JFA framework, *D* is initialized as a random vector and estimated with or without *V*. Outside of the JFA framework, D can be estimated as$D^2 = {}^{\Sigma_0}/_{\tau}$; τ is the regulation factor that controls the prior distribution for maximum a posteriori (MAP) adaptation.

Table 1 shows results for this system on 2008 SRE short2-short3 conditions 7 and 6. The table's first row ("NO") shows the effect of no adaptation (without D), conventional MAP (with D=UBM), and D estimated using JFA statistics (D=JFA). Results without D are obtained with supervectors that are maximum likelihood estimates given the features and the UBM. Although D from JFA is not used, there is an implicit uniform prior with ML estimate. Results within a row are comparable to each other, and the best results are obtained with D estimation from JFA.

The second row ("CHANNEL") results obtained using an eigenchannel approach where 300 channel factors are removed. The resulting supervector is estimated as $m - Ux$. The results are very little with the use and estimation of D. The third row ("SPEAKER") shows the results obtained with the eigenspeaker approach where the supervector is estimated as$m_0 + Vy$. They depend heavily on the use of D. The results improve with D and improve further with D estimated from the UBM. The fourth row ("CHANNEL,SPEAKER") shows results obtained with JFA model. They show a trend similar to that of the third row ("SPEAKER") except that D=UBM gives the best performance for English trials and D=JFA gives the best performance for all trials.

The performance on all trials can be significantly improved by language compensation [7], which is not used in this paper for two reasons. First, it requires another database, such as 2006 SRE, for training the compensation, but there is significant mismatch between 2006 and 2008 SRE results in terms of the effect of JFA. Second, the improvement in performance on all trials may result from better calibration across different language conditions in addition to better performance for each condition. Language compensation masks the former improvement.

4. ACROSS-PHONE VARIABILITY

Across-phone variability is computed as shown in Table 2. For each conversation side, feature vectors after mean variance normalization are selected based on the phone label. As shown in the table, we use four phone classes: Vowels, Glides+Nasals, Obstruents, and Pause. Ideally, Pause could be ignored, but it is included because the frames labeled as pause by the open-phone loop phone recognizer after speech/silence segmentation will belong to unvoiced phones and may be useful for speaker recognition. In addition, we want to make sure that combining all the frames from all the classes gives all the frames used by the baseline system. The GMM supervector is estimated with frames for each class. A given speaker (with a given channel) supervector can be described as a sum of UBM mean m_0 and phone supervector Pq

$$m = m_0 + Pq$$

The phone supervector distribution lies in a low-dimensional subspace of rank R_P and is assumed to be distributed according to PP^T. The phone factors q are assumed to have a standard normal distribution. The phone subspace P is estimated from speaker, channel and phone specific supervectors in the same way as the nuisance subspace in the conventional JFA.

Table 1 %EER for baseline JFA system with three different choices of D evaluated on 2008 SRE short-2short3 conditions 7 (ENG) and 6 (ALL). Bold cell indicates that the subspace was removed. Resulting supervector dimensions are indicated within brackets. If D is not used then the dimensions are as specified in Column 1. If D is used then the dimensions are NF. First row ("NO") and first column ("WITHOUT D") result uses maximum likelihood estimate of the supervector using UBM and the data.

JFA (Supervector Dimensions)	WITHOUT D		WITH D=UBM (*NF*)		WITH D=JFA (*NF*)	
	ENG	ALL	ENG	ALL	ENG	ALL
NO (*NF*)	9.039	12.920	8.958	12.509	8.632	12.509
CHANNEL (*NF*-300)	**3.176**	**7.618**	**2.932**	**7.580**	**3.095**	**7.655**
SPEAKER (300)	12.866	16.094	10.831	14.264	12.296	15.571
CHANNEL, SPEAKER (300)	3.990	7.169	2.769	7.020	3.420	6.647

Table 2 Setup for computing across-phone variability

Speaker, Channel	Phones			
s1,1	s1,1 Vowel	s1,1 GN	s1,1 Obstr	s1,1 Pause
s1,2	s1,2 Vowel	s1,2 GN	s1,2 Obstr	s1,2 Pause
s2,1	s2,1 Vowel	s2,1 GN	s2,1 Obstr	s2,1 Pause
.	.	.	.	.
sN,M	sN,M Vowel	sN,M GN	sN,M Obstr	sN,M Pause

5. RESULTS WITH FOUR CLASSES

We estimate channel, speaker, and phone subspaces independently using the NIST 2004 SRE dataset. We use 300 dimensions for each subspace. The goal of this paper is to compare the performance of these subspaces. As mentioned, we will compare the results without diagonal factors (D) first to compare only the effect of subspaces. We will concatenate the subspaces as channel+phone and speaker+phone and measure improvement over individual subspaces. We will perform experiments in the JFA framework with speaker and channel subspaces where the phone subspace is added to one of them.

5.1.1. Results with individual phone, speaker and channel subspaces

Table 3 %EER obtained by removing channel, speaker and phone subspaces

Remove Subspace (*NF-300 dim*)		Without D	
		Eng trials	All trials
1	None (*NF*)	9.039	12.920
2	Channel	3.176	7.618
3	Speaker	5.945	10.754
4	Phone	6.922	10.866

Table 4 %EER obtained by keeping channel, speaker and phone subspaces

Keep Subspace (300 dim)		Without D	
		Eng trials	All trials
1	All (*NF*)	9.039	12.920
2	Channel	15.228	18.297
3	Speaker	12.866	16.094
4	Phone	14.169	17.140

Table 3 and Table 4 show the effects of individual subspaces being used or removed. The idea is to compare the results with the baseline. If a particular subspace represents nuisance dimensions, then removing it will improve the performance and using only the subspace will worsen the performance. This hypothesis is supported only by the channel subspace. Removing speaker (and phone) subspaces shows the same trend as channel. Performance improves by removing the subspaces and worsens by keeping them. This is counter intuitive because speaker subspace contains useful variability so removing it should hurt the performance. This results shows that speaker and channel subspaces do not represent speaker and channel information separately, but contain both. This has been reported in [8], where Dehak et al. showed that performance with speaker factors can be improved by within-class covariance normalization [9], which is a channel compensation technique. Overall, the best subspace to keep seems to be speaker and the best subspace to remove seems to be channel.

The performance of the phone subspace shows that it has the least effect when it is removed, which implies that either it is relatively small or it is not highly correlated with the channel subspace. The latter is a very interesting hypothesis because the phone subspace is contained in the channel subspace. If the hypothesis is correct, it may point to a correlation between speaker and channel subspaces.

5.1.2. Results with combination of phone, speaker and channel subspaces

We now explore the similarities between the subspaces by combining them. Table 5 shows the results of these experiments. Row 1 is the baseline showing that removing the channel subspace and keeping the speaker subspace results in better performance on all trials but worse performance on English trials compared to just removing the channel subspace (Table 3, row 2). First we test the hypothesis that the phone subspace contains mostly speaker or channel variability. In that case, performance will not degrade much if we replace the speaker or channel subspaces with the phone subspace. Rows 2 and 3 show that the performance significantly degrades when replacing speaker and channel subspaces in baseline with the phone subspace. In both cases, the results do not support the hypothesis. Next we test the hypothesis that phone subspace contains new information for speaker recognition which is not modeled by existing speaker and channel subspaces. Therefore, better performance can be obtained by combining the phone subspace with the speaker or channel subspace. Row 4 shows the improvement in performance on English trials when the phone subspace is appended to the speaker subspace. This is an interesting result because the phone subspace is a part of the channel subspace. It is also interesting because the results show improvement for English trials when the alignments are obtained with a language-independent open phone loop phone recognizer. Row 5 shows that the performance is not affected when the phone subspace is appended to the channel subspace. This confirms our earlier hypothesis about the nature of these subspaces.

Table 5 %EER with concatenated subspaces in JFA framework

Keep Subspace		Remove Subspace	Without D	
			Eng trials	All trials
1	Speaker	Channel	3.990	7.169
2	Phone	Channel	7.166	11.277
3	Speaker	Phone	11.319	14.339
4	Speaker + Phone	Channel	3.583	7.132
5	Speaker	Channel + Phone	3.827	7.319

Note that Table 3 and Table 5 show results without the diagonal term. In Table 6 and Table 7, the diagonal term is added to the JFA framework. As mentioned earlier, there are two ways of obtaining the diagonal term – one using JFA and other using UBM. As seen in Table 1, the choice of D has a significant effect on performance.

Table 6 shows that when D is estimated from JFA, the trend is similar to Table 4. The performance improves when phone and speaker subspaces are concatenated for English trials. The performance is not affected when phone and channel subspaces are concatenated. Note that D was estimated with only speaker and channel factors. Our experiments show that similar performance is obtained when D is reestimated for every configuration.

Table 6 %EER with channel, phone and speaker subspaces in JFA framework with D estimated from JFA

Keep Subspace	Remove Subspace	D from JFA	
		Eng	All
Speaker	Channel	3.420	6.647
Speaker+Phone	Channel	3.095	6.908
Speaker	Channel+Phone	3.339	6.796

Table 7 shows that when D is estimated from the UBM, the trend is very different. There is no improvement after concatenating the phone subspace with the speaker or channel subspace. It can be hypothesized that the diagonal term from the UBM contains phone information similar to that in the phone subspace. This is an interesting hypothesis because D is not explicitly trained for phone variability. More work is needed to test this hypothesis.

Table 7 %EER with channel, phone and speaker subspaces in JFA framework with D estimated from UBM

Keep Subspace	Remove Subspace	D from JFA	
		Eng	All
Speaker	Channel	2.769	7.020
Speaker+Phone	Channel	2.769	7.095
Speaker	Channel+Phone	2.850	6.946

6. SUMMARY

We estimate the usefulness of across-phone variability for speaker recognition in JFA framework. We used a Hungarian open-phone loop recognizer to estimate this variability. The variability is compared to speaker and channel variability. All the variabilities were estimated as 300 dimensional subspaces in an about-60k dimensional mean supervector space using JFA. First we compared the subspaces by keeping and removing them. Results showed that the most harmful subspace is channel and the most useful subspace is speaker. Result also showed that better performance is obtained by removing the speaker subspace and performance using only the channel subspace is not much worse than the performance using only the speaker subspace. This supports previous observations that speaker and channel subspaces contain both types of variability. The results obtained by removing phone subspace produced the least degradation and those obtained by keeping phone subspace were in between speaker and channel.

These results were not conclusive to determine the usefulness of phone variability for speaker recognition. To that end, we ran experiments by concatenating the subspaces. This is an approximation to the joint estimation. The baseline is a JFA system where channel subspace is removed and speaker subspace is preserved. The results showed that the performance does not change when phone subspace is concatenated with channel subspace and is removed. This shows that the phone subspaces does not contain more nuisance dimensions than channel subspace. However the results showed improvement when phone subspace was concatenated with speaker subspace and was preserved. This shows that phone subspace contains useful variability for speaker recognition.

Note that the results obtained so far did not use the diagonal term from JFA. We ran the same experiments with different ways of estimating the diagonal term. The improvement persisted when the diagonal term was estimated from JFA data but was reduced with the diagonal term estimated from the UBM. This shows that there is an interaction between phone variability and the variability represented by the diagonal term. More work is needed to understand the interaction.

7. DISCUSSION AND FUTURE WORK

Our work has several interesting or surprising results. The phone subspace is estimated using a language-independent phone recognizer but its use only improves the performance of English trials. In one case, it worsens the performance over all trials.

The phone subspace is a part of the conventional channel subspace but it seems to combine well with the speaker subspace. Therefore, future work will explore possible ways of estimating the phone subspace with the speaker and channel subspaces. For example, the phone subspace can be estimated first, followed by the channel and speaker subspaces. Removing the phone subspace might improve the estimation of the channel subspace. Phone factors can be used with short test utterances similar to [2].

We will replicate these results with the SRI English ASR system and compare results with those obtained from BUT's open phone loop recognizer. The choice of classes is important for the results and it should be investigated whether other choices of four or more classes can produce similar improvements. We will rerun these experiments with an improved baseline to observe the effect of datasets and features on the conclusions.

8. ACKNOWLEDGMENTS

The author thanks Lukas Burget and the Brno University of Technology team for organizing the 2008 JHU summer workshop and for generously providing the open phone loop output. The author also thanks Nicolas Scheffer for numerous discussions on JFA and for many helpful hints. This work was funded through a development contract with Sandia National Laboratories (#DE-AC04-94AL85000). The views herein are those of the author and do not necessarily represent the views of the funding agencies.

9. REFERENCES

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," Proc. of ICASSP, vol. 1, pp. 637-640, Philadelphia, PA, 2005.

[2] R. Vogt, C. J. Lustri, and S. Shridharan, "Factor analysis modeling for speaker verification with short utterances," Proc. of Odyssey 2008: The Speaker and Language Recognition Workshop, Stellenbosch, South Africa, 2008.

[3] S. Kajarekar, "Phone-based cepstral polynomial SVM system for speaker recognition," Proc. of Interspeech, Brisbane, Australia, 2008.

[4] N. Scheffer, R. Vogt, J. Pelecanos, and S. Kajarekar, "Combination strategies for the JFA model in speaker verification, application to a phonetic system," Proc. of ICASSP, Taipei, Taiwan, 2009.

[5] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," Proc. of ICASSP, Taipei, Taiwan, 2009.

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

[7] S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, "The SRI NIST 2008 speaker recognition evaluation system," Proc. of ICASSP, Taipei, Taiwan, 2009.

[8] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," Proc. of ICASSP, Taipei, Taiwan, 2009.

[9] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," Proc. of ICSLP, pp. 1471-1474, Pittsburgh, PA, 2006.