INTERSPEECH 2009 BRIGHTON

Within-Session Variability Modelling for Factor Analysis Speaker Verification

Robbie Vogt[1], *Jason Pelecanos*[2], *Nicolas Scheffer*[3], *Sachin Kajarekar*[3], *Sridha Sridharan*[1]

[1]Speech Research Lab, Queensland University of Technology, Brisbane, Australia.
[2]IBM T.J. Watson Research Center, Yorktown Heights, NY.
[3]SRI International, Menlo Park, CA.

{r.vogt,s.sridharan}@qut.edu.au, jwpeleca@us.ibm.com, {scheffer,sachin}@speech.sri.com

10.21437/Interspeech.2009-386

Abstract

This work presents an extended Joint Factor Analysis model including explicit modelling of unwanted within-session variability. The goals of the proposed extended JFA model are to improve verification performance with short utterances by compensating for the effects of limited or imbalanced phonetic coverage, and to produce a flexible JFA model that is effective over a wide range of utterance lengths without adjusting model parameters such as retraining session subspaces. Experimental results on the 2006 NIST SRE corpus demonstrate the flexibility of the proposed model by providing competitive results over a wide range of utterance lengths without retraining and also yielding modest improvements in a number of conditions over current state-of-the-art.

Index Terms: speaker recognition, factor analysis, within-session variability.

1. Introduction

The introduction of Joint Factor Analysis [1] has seen widespread adoption in the speaker verification community due to the large reduction in error rates achieved in recent NIST Speaker Recognition Evaluations (SRE). These improved error rates have been achieved largely through the substantial modelling power provided by the JFA architecture, particularly in modelling differences between recordings of the same speaker (inter-session variability).

Recent observations have shown that the current Joint Factor Analysis (JFA) model does not provide the expected improvements in performance for short utterance lengths that it does for the core NIST SRE condition using full conversation sides [2]. It is hypothesised in this work that this poor performance is the result of deficiencies in the current JFA model particularly with respect to modelling the unwanted variability present within the session.

Based on these observations, an extended JFA model is introduced in this work to specifically address the characteristics of verification with short utterances by incorporating explicit modelling of within-session variability, such as the phonetic information encoded in an utterance. The goals of this extended JFA model are specifically to improve speaker verification performance with short utterances such as in the range 5–20 seconds of active speech, and also to develop a model that generalises well to a wide range of utterance lengths without need for retraining.

The following section investigates the effect of verification with short utterances using the standard JFA approach, through the results of recent studies, highlighting the role of session variability and its dependency on utterance length. Section 3 then proposes the extended factor analysis model that incorporates within-session variability modelling to combat the deficiencies of the standard model. Implementation details and experiments on NIST SRE 2006 data are then presented with a brief discussion in Sections 4 and 5, respectively. Finally, a summary and possible future directions are presented in Section 6.

2. Deficiencies of JFA for Short Utterances

Previous work has highlighted some deficiencies with current Joint Factor Analysis models for shorter utterance lengths. Particularly, the effectiveness of JFA, and session variability modelling in particular, is reduced with shorter utterances for training and testing as demonstrated by the results in [2]. While JFA provides a significant performance improvement for full conversation sides, this improvement diminishes with utterance lengths restricted to 20 seconds or less for training and testing. Speaker factors were found to be generally beneficial, but the inclusion of session factors had a significant *negative* impact on performance when utterance lengths were restricted to 20 seconds or less for training and testing.

Further investigation in [3] found that training the session variability subspace matrix $\boldsymbol{U}$ with utterances of matched length to the evaluation conditions resulted in performance gains using the full JFA model with session factors even with utterance lengths as short as 10 and 20 seconds. From these results it is concluded that the observed variability between sessions is dependent on utterance length. The improved performance attained with matched session subspaces in [3] indicates that the matched session subspaces are substantially different at different utterance lengths.

This observed behaviour does not fit well with the assumptions made by the JFA model. It has been assumed to this point that the session factors and session subspace capture environmental effects such as channel, handset and background noise as this was the initial intent [1, 4]. The characteristics of these environmental effects should have consistent characteristics regardless of utterance length.

This dependency on utterance length is problematic firstly because we are thus required to train specialised session subspaces for the range of utterance lengths of interest to extract optimal performance from the JFA model, but, more importantly, it implies that our assumptions about the nature of the inter-session variability are flawed.

It was noted in [3] that shorter utterances show an *increase* in overall session variability as measured by the trace of the session subspaces for various utterance lengths (Table 4 in [3]). A reasonable explanation for this result is that the consistent, stationary environmental factors may well still be present as ut-

terances become shorter, but an additional source of variability is becoming more apparent with reducing utterance length.

One hypothesised source of this extra captured variability is the variability introduced by the speech content, that is the phonetic information encoded in the speech. This is unwanted variability in the context of text-independent speaker recognition. For typical NIST conversation lengths, there is likely to be a reasonable coverage of the phonetic space and the effects of phonetic variability will largely average out. For utterances of only a few seconds in length, however, there will be very poor coverage of the phonetic space, and differences in the particular observed phones have the potential to result in significant biases in the produced speaker model estimate.

3. Modelling Within-Session Variability

This work extends the current JFA model. The goals of extending the model are two-fold:

1. To produce better performance from the JFA model in the specific case of short utterances, through using a model that better fits the underlying process.
2. To construct a JFA model that is effectively *independent of utterance length* in order to avoid the issues of retraining the session subspace for different length training and testing utterances. This is particularly relevant if an evaluation or application has mixed utterance lengths or the utterance length is not known a priori.

With these goals in mind, the approach taken in this work is to extend the JFA model by separating the sources of session variability—a collective term for all information *not* useful for identifying a speaker—into distinct and independent sources of *inter*-session variability and *within*-session variability. Central to the idea of observing and modelling within-session variability with the JFA model is the idea of dividing an utterance into a sequence of N short segments each described by a GMM mean supervector $\boldsymbol{s}_n; n = 1, \ldots, N$. While all of the short segment means in an utterance are constrained to have the same speaker and inter-session characteristics, they are permitted to vary in a very low dimensional within-session variability subspace. The complete FA model for a short segment n of an utterance is therefore

$$\boldsymbol{s}_n = \boldsymbol{m}_s + \boldsymbol{U}_I\boldsymbol{x} + \boldsymbol{U}_W\boldsymbol{w}_n,$$

where $\boldsymbol{m}_s$ is the usual JFA speaker mean [1] given by

$$\boldsymbol{m}_s = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z}.$$

The reader is referred to [1] for a detailed description of the standard JFA model. While $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are all held constant for an utterance, there will be independent within-session factors, $\boldsymbol{w}_n$, for each short segment n.

In this extended model, inter-session variability is modelled as an offset $\boldsymbol{U}_I\boldsymbol{x}$ to the GMM mean supervector. That is, $\boldsymbol{U}_I$ is equivalent to $\boldsymbol{U}$ in the standard JFA model, except that $\boldsymbol{U}_I\boldsymbol{x}$ is intended to strictly represent only constant environmental effects such as handset and channel. A goal therefore is to train $\boldsymbol{U}_I$ in such a way as to capture only stationary environmental effects that are independent of utterance length.

Additionally, within-session variability is modelled over a shorter time span than the inter-session variability to capture and remove transient effects within an utterance.

Following the hypothesis that phonetic variability is the dominant source of this transient, within-session variability, this work explores modelling within-session variability for short segments that are aligned with open-loop phone recogniser (OLPR) transcripts. Using this alignment, each phone instance in the OLPR transcript is mapped to a short segment (only the start- and end-times from the phone instances are retained, while the phone labels are disregarded).

The OLPR transcripts are derived from the BUT Hungarian phone recognition system [5]. This phone recogniser has previously been shown to be effective for a number of applications including speech activity detection and language recognition. On the Mixer date used in recente SRE's, this recogniser produces phone events with an average length of 10 speech frames.

There are other potential methods of segmenting an utterance which may deserve pursuing. One potential option is to simply segment the active speech of an utterance at regular intervals, say 0.1–1 seconds of active speech. This has the advantage of not requiring a phone recogniser, and a consistent segment length may result in more consistent $\boldsymbol{w}_n$ estimates. Another possibility is aligning segments with syllables, allowing the segments to be centred around high-energy syllable nuclei. This may provide better quality $\boldsymbol{w}_n$ estimates due to cepstral representations of high-energy voiced speech generally being less affected by environmental effects but would obviously require syllable transcripts or some method of recognising syllable-like event boundaries, such as used in [6].

4. Implementation

Several systems were developed for comparison in this work. Details of these systems are presented below.

4.1. Baseline JFA System

The baseline system for this evaluation implemented the standard JFA model introduced by Kenny, *et al.* [1] for speaker modelling. This implementation was based on a "small" system comprising 512-component, gender-independent GMM's with 39-dimensional MFCC-based features. Details of the features and UBM training data are given in [7].

For simplicity and efficiency, this system implemented dot product scoring for the verification trials as described in [8]. Channel compensation was also applied to the statistics for both training and testing utterances in the manner described in [8]. Gender-dependent ZT-norm was also applied using around 300 utterances per gender.

Gender-independent JFA parameters for the baseline system were trained on a relatively small subset of Mixer data, more specifically, $\boldsymbol{U}$ and $\boldsymbol{V}$ were trained on SRE 04 utterances from around 300 speakers while $\boldsymbol{d}$ was trained on a disjoint set of utterances from 57 speakers mostly from SRE 05. The subspace dimensions were limited to 100 speaker factors and 50 session factors. This baseline configuration was chosen to be representative of a larger state-of-the-art JFA system while requiring a dramatically reduced computational load.

4.2. Matching U to the Utterance Length

This system is identical to the baseline system in most aspects except in the data used to train the session subspace transform $\boldsymbol{U}$. Additional $\boldsymbol{U}$ matrices were trained using the same utterances as the baseline system, except the utterances were truncated in length to match the anticipated utterance lengths to be used in the experiments. In this way, additional $\boldsymbol{U}$ matrices specialised for 20-second and 10-second conditions were produced for this system, while $\boldsymbol{V}$ and $\boldsymbol{d}$ were unchanged. The choice of specialised session transform is made independently for training

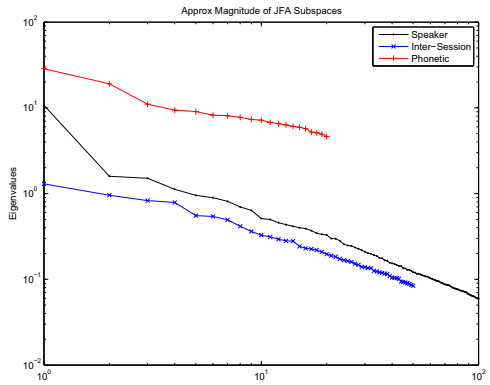

Figure 1: *Leading eigenvalues of the speaker, inter-session and within-session variability. The within-session subspace was trained on segments aligned to OLPR transcripts.*

and testing, therefore when the training length is different to the testing length different $\boldsymbol{U}$ matrices are used between training and testing (eg. whole conversation training, 10-second testing).

4.3. Extended JFA System Incorporating Within-Session Modelling

The extended JFA system in these experiments was developed to be as similar to the baseline system as possible. The parameters of the standard part of the JFA model ($\boldsymbol{U}_I$, $\boldsymbol{V}$ and $\boldsymbol{d}$) were identical to those used for the baseline system while the additional within-session subspace, $\boldsymbol{U}_W$, was trained on a subset of approximately 100 utterances (2 from each of the 50 speakers). This training process is analogous to the training of $\boldsymbol{U}_I$, except that $\boldsymbol{U}_W$ is trained to capture the dominant directions of differences between the short segments of the training utterances. The transcripts from the Hungarian OLPR [5] provided the segment alignment for the utterances used in estimating $\boldsymbol{U}_W$. It is therefore expected that the within-session variability captured through this procedure will be dominated by phonetic information, however, it is also reasonable to expect variation in actual realisations of phones will also be present.

The leading eigenvalues of the speaker, inter-session and within-session variability resulting from this training procedure are plotted in Fig. 1. The raw within-session variability is evidently very high; approximately an order of magnitude greater than both the speaker and inter-session variability. The *effective* within-session variability over the length of an utterance, though, will be substantially less. For utterances of around 1–10 seconds (or 10–100 phone events) within- and inter-session variability will be of similar magnitudes. As the utterance length increases to a full conversation side of approximately 100 seconds, the effect of within-session variability to the utterance as a whole will be effectively negligible, as expected, due to the averaging effect and sufficient coverage of the phonetic space.

During both speaker model training and testing, the effect of within-session variability was removed in an analogous fashion to the inter-session variability of the baseline system: For each short segment n of an utterance, the within-session factors $\boldsymbol{w}_n$ were estimated and the sufficient statistics compensated as in [8]. The compensated statistics for all of the segments in an utterance were then summed to give utterance statistics with within-session effects removed. These within-session-compensated statistics were then used in the same way as the usual utterance statistics in the baseline system.

5. Experiments

A system implementing the extended JFA model with within-session variability was evaluated and compared against the standard and matched-$\boldsymbol{U}$ JFA systems on the NIST SRE 2006 English-only, *1conv4w-1conv4w* condition. To investigate the performance of the systems with reduced utterance lengths, the *1conv4w-1conv4w* condition was again utilised however both the training and testing utterances were truncated to produce shorter utterances of 10 and 20 seconds. From the results in [2] and [3], the 10- to 20-second utterance length range appears to be the range at which the effectiveness of the standard JFA model is diminishing.

Table 1 presents EER and minimum DCF results comparing the variants of the JFA model. The first and second rows use the standard JFA model with 50 and 60 session factors, respectively. The third row shows the results with $\boldsymbol{U}$ matched to the length of utterance used for training and testing. The last row includes within-session variability modelling with 50 inter-session factors and 10 within-session factors.

As reported in [3], matching $\boldsymbol{U}$ to the evaluation conditions provides an advantage over the standard JFA model. The matched system provided better performance in all short conditions over the baseline although the improvement for the 10-sec condition is quite modest.

Incorporating within-session variability modelling largely produced similar results to the matched-$\boldsymbol{U}$ approach, improving on the standard JFA system for all shortened utterances. Additionally, at the EER operating point this approach gave the best performance at each utterance length, although only by a small margin. Results were less clear-cut when measured by minimum DCF.

From these results it can be seen that the introduction of within-session factors at least achieved one of the stated goals of producing a system that could be effective over a wide range of utterance lengths. While the matched system used a distinct $\boldsymbol{U}$ matrix for each utterance length tested, the parameters of the within-session modelling system were consistent across all trials. Thus, the within-session modelling approach provides a practical advantage over the standard JFA model through its flexibility.

The second goal of improving performance through more accurately modelling the unwanted variability has not been convincingly achieved with these results. Several factors may contribute to this outcome, such as the optimal choice of segmentation, but more importantly the approach to estimating the subspaces of the extended model used for these experiments was not at all tailored to the extended model. The effects of including within-session modelling on the speaker and inter-session subspaces are likely to be substantial. Future investigation of segmentation choice and proper integration of within-session modelling in the subspace estimation process may lead to significant improvements in performance of this extended model.

An added complication is introduced when the training and testing utterance lengths differ. In this case, the "matched" matrix $\boldsymbol{U}$ is different for training and testing. Table 2 presents results evaluated with a whole conversation for training but only 20 or 10 second testing utterances.

Again in this table the first row is the baseline approach using the standard JFA model. The results in the second represent a system with $\boldsymbol{U}$ independently matched to the utterance length for training and the utterance length for testing. Interestingly, while the matched-$\boldsymbol{U}$ approach worked quite well with the same utterance lengths for both training and testing, it causes a degra-

Table 1: *Comparison of EER and minimum DCF performance for the standard JFA model, matched-length session JFA model and the extended JFA model incorporating within-session variability modelling on the SRE 06 common evaluation condition with truncated utterances for both training and testing.*

JFA Model	Dims	1 conv		20 sec		10 sec	
$\boldsymbol{U}$	50	3.10%	.0159	12.79%	.0561	20.21%	.0819
$\boldsymbol{U}$	60	3.03%	**.0156**	13.01%	.0562	20.31%	.0820
$\boldsymbol{U}_{\text{Matched}}$	50	3.10%	.0159	12.20%	**.0531**	19.71%	.0814
$\boldsymbol{U}_I + \boldsymbol{U}_W$	$50_I + 10_W$	**2.97%**	.0170	**11.98%**	.0541	**19.67%**	**.0807**

Table 2: *Comparison of EER and minimum DCF performance for the standard JFA model, matched-length session JFA model, a stacked session model and the extended JFA model incorporating within-session variability modelling on the SRE 06 common evaluation condition with whole conversation side training and truncated utterances for testing.*

JFA Model	Dims	20 sec		10 sec	
$\boldsymbol{U}$	50	6.12%	.0293	9.59%	.0433
$\boldsymbol{U}_{\text{Matched}}$	50	6.39%	.0305	10.13%	.044
$\boldsymbol{U}_{\text{Stacked}}$	100	5.91%	**.0275**	**9.54%**	.0421
$\boldsymbol{U}_I + \boldsymbol{U}_W$	$50_I + 10_W$	**5.85%**	.0290	9.59%	**.0414**

dation in performance in all measures compared to the baseline system.

The third row of Table 2 demonstrates that a stacking approach [9] provides an improvement in all cases over the baseline system, regaining the modest advantage of the matched approach observed previously. Under the stacking approach, a larger session subspace was constructed by concatenating the two session matrices matched to the training and testing conditions, for example, for a 1conv training, 10 second test condition, the $\boldsymbol{U}$ used for both training and testing consists of concatenated matrices matched to the 1 conv and 10 second utterance lengths. This approach has been successfully employed previously for mixed telephone and distant microphone conditions in recent SRE's in 2006 and 2008.

Finally, the last row in Table 2 presents the performance of incorporating within-session modelling. As with the stacking approach, the extended model provides improved performance over the baseline system in all cases, except for the 10-sec EER where the two are equivalent. The extended approach is also competitive with the stacked approach as they each provide the best performance depending on the condition and performance measure.

The results for these experiments again highlight the ability for the extended JFA model to provide competitive performance across a wide range of operating conditions without having to adjust model parameters. This flexibility is a major advantage of this approach, especially for situations in which it is not possible to know the training and testing utterance lengths prior to evaluation or, as in this case, the utterance lengths are not consistent for training and testing.

6. Summary

Motivated by relatively poor and ineffective performance for short utterance lengths, this work presented an extension to the joint factor analysis model to include modelling of unwanted within-session variability. The inclusion of within-session variability modelling was particularly intended to compensate for the effects of uneven phonetic coverage for short utterances by modelling and removing the effects of phonetic variation over short segments of each utterance.

The goals of the extended model were to produce better performance from the JFA model in the specific case of short utterances by using a more realistic model, and to produce a flexible JFA model that would be effective over a wide range of utterance lengths without adjusting model parameters such as retraining session subspaces.

Experimental results demonstrate the flexibility of the extended JFA model by providing competitive results over a wide range of utterance lengths and operating conditions without need for adjusting any of the model parameters. While modest performance improvements were also observed in a number of conditions over current state-of-the-art, further work is necessary to demonstrate that significant performance improvements are achievable through this extended model.

7. Acknowledgments

The authors thank the JHU for the 2008 Summer Workshop as this work was accomplished during that time, as well as the BUT Speech Group for the contribution of their state-of-the-art JFA system, computing resources during and support during the Workshop. The authors would also like to thank and acknowledge the contribution of Elly (Hye Young) Na to this work.

The research by authors at QUT was supported by the Australian Research Council Discovery Grant No. DP0877835.

8. References

[1] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 219–226.

[2] R. Vogt, C. Lustri, and S. Sridharan, "Factor analysis modelling for speaker verification with short utterances," in *Odyssey: The Speaker and Language Recognition Workshop*, 2008.

[3] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech*, 2008, pp. 853–856.

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

[5] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," in *International Conference on Text, Speech and Dialogue*, 2004, pp. 465–472.

[6] N. Dehak, P. Kenny, and P. Dumouchel, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Trans. on ASLP*, vol. 15, pp. 2095–2103, 2007.

[7] L. Burget, P. Matějka, P. Schwarz, O. Glembek, and J. Černocký, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. on ASLP*, vol. 15, no. 7, pp. 1979–1986, Sept. 2007.

[8] A. Strasheim and N. Brümmer, "SUNSDV system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop Booklet*, 2008.

[9] N. Scheffer, R. Vogt, S. Kajarekar, and J. Pelecanos, "Combination Strategies for a Factor Analysis Phone-Conditioned Speaker Verification System," in *ICASSP*, 2009.