A COMPARISON OF APPROACHES FOR MODELING PROSODIC FEATURES IN SPEAKER RECOGNITION

Luciana Ferrer[*] *Nicolas Scheffer* *Elizabeth Shriberg*

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA

ABSTRACT

Prosodic information has been successfully used for speaker recognition for more than a decade. The best-performing prosodic system to date has been one based on features extracted over syllables obtained automatically from speech recognition output. The features are then transformed using a Fisher kernel, and speaker models are trained using support vector machines (SVMs). Recently, a simpler version of these features, based on pseudo-syllables was shown to perform well when modeled using joint factor analysis (JFA). In this work, we study the two modeling techniques for the simpler set of features. We show that, for these features, a combination of JFA systems for different sequence lengths greatly outperforms both original modeling methods. Furthermore, we show that the combination of both methods gives significant improvements over the best single system. Overall, a performance improvement of 30% in the detection cost function (DCF) with respect to the two previously published methods is achieved using very simple strategies.

***Index Terms*—** Speaker recognition, Prosody, Joint Factor Analysis, Support Vector Machines

1. INTRODUCTION

We consider the task of text-independent speaker verification: given a sample from a speaker and a claimed identity we need to decide whether the claim is true or false. A successful approach to speaker verification is to combine different knowledge sources by modeling them separately, fusing them at the score level to produce the final score that is later thresholded to obtain a decision. Combinations of systems are most successful when the individual systems being combined are significantly different from each other. Currently, the best individual speaker recognition systems are based on low-level spectral features modeled using joint factor analysis (JFA) techniques. Prosody, the intonation, rhythm and stress patterns in speech, is not directly reflected in the spectral features and, hence, a system based on prosodic information should be highly independent from a low-level spectral system. Indeed, it has been shown [1, 2, 3, 4] that systems based on prosodic information can lead to significant improvements when combined with a state-of-the-art low-level system.

Several speaker recognition systems have been proposed in the last decade [5, 6, 7, 3]. All these systems define regions of extraction based on some event in the waveform that can be determined automatically and then extract certain measurements based on the pitch and energy signals and, sometimes, the durations of subregions within them. Several methods have been used to model these features. In this paper we will focus on comparing two general modeling methods: one that models Gaussian mixture weights using support vector machines (SVMs) [8] and the currently standard method for modeling spectral features: JFA [9, 3]. We perform the study on a small subset of well-behaved prosodic features used in [3] that can be modeled using current JFA techniques. Furthermore, we propose three extensions to the JFA system presented in that paper: (1) the use of several degrees for the Legendre polynomial approximation used to compute the features, (2) the modeling of sequences of consecutive feature vectors to capture their dynamic behavior, and (3) the combination of the JFA system with the GMM-SVM system. Overall, we show gains in DCF of up to 30% with respect to the methods previously presented in the literature.

2. PROSODIC FEATURES

In [7] we presented a paradigm for the extraction of prosodic features from speech. Syllables are estimated automatically using the output of an automatic speech recognition (ASR) system, and more than a hundred measurements based on pitch and energy signals, along with the duration of the syllable and its constituents (onset, nucleus, and coda) are extracted over each syllable. We called these features syllable-based NERFs (non-uniform extraction region features), or SNERFs. The extracted features have some particular characteristics that make them harder to model than the standard spectral features: they have mixed continuous/discrete distributions, they are much sparser than low-level features, and they have undefined values. A system based on these features has been the best performing prosodic system on NIST speaker recognition evaluation (SRE) data published in the speaker recognition literature, since its introduction in 2005.

Despite its success, these syllable-based features have not been widely used in the community, probably because they are not simple to extract. They require ASR output and, even though they are all basically simple measurements over the pitch, energy and duration patterns, their implementation is laborious. In [3], Dehak proposed the use of ASR-independent regions based on the valleys found in the energy signal and the use of polynomial approximations of the pitch and energy signals, along with the length of the regions, as features. Furthermore, they proposed the use of JFA for their modeling.

In this paper our focus is on comparing modeling methods for the simpler set of features, which we will call *energy valley-based polynomial approximation* (EV-PA). A detailed description of the extraction procedure for these features can be found in [10]. Essentially, the speech signal is segmented into regions by splitting the voiced regions wherever the energy signal reaches a local minimum. The minimum is obtained in our case by finding the positive crossings through zero of the derivative signal, estimated using a smooth delta function given by $\delta(i) = 0.10(x(i+1) - x(i-1) + 2x(i+2) - 2x(i-2))$, where $x(i)$ is the energy at frame i. The energy and pitch signals are obtained using the get_f0 function from the Snack toolkit [11]. For each region, these signals are approximated with a Legendre polynomial of order N (set as 5 in the original work from

[*]This author performed part of the work presented in this paper while at the Information Systems Laboratory, Department of Electrical Engineering, Stanford University.

Dehak et al.). The length of the region is also used as a feature. A total of $2N + 1$ features is then extracted for each EV region.

It is important to note that the EV-PA features are not just simpler than our original set of syllable-based features in that they do not require ASR and they are easier to extract. They are also simpler to model since they are all continuous and do not contain undefined values. This makes the JFA modeling of these features possible and is why we have chosen them to perform the current study instead of using the original set of features. The generalization of the JFA method to the more general set of prosodic features is one of our current priorities and one that, all current evidence indicates, should result in significant improvements.

3. SVM MODELING OF GMM WEIGHTS

As mentioned in the previous section, the SNERFs cannot be simply modeled using Gaussian mixture models (GMMs) since the features contain undefined values. The first attempt at modeling these features was to adapt the GMM-UBM (Gaussian mixture model - universal background model) modeling method [12] by adding a probability of undefined value to each Gaussian in the mixture [13]. Another approach is to do SVM modeling of some parameterization of the distribution of these features [8]. This is the approach we use for this paper, since it was shown to give good results on the latest set of SNERFs. The general method is to train a UBM on held-out data, given by a GMM with the additional probabilities of undefined value, as in the original GMM-UBM method. The GMM weights are then adapted to each sample, and the vector of adapted weights is used as a feature vector that is then modeled using SVMs.

Since the dimensionality of the feature vector is large and the features are sparse, a back-off strategy is used: several UBMs are trained for different subsets of features instead of a single UBM for the complete set. In this work we train UBMs for each individual feature, and for groups of all same-order polynomial coefficients. Furthermore, we have found that sequences of prosodic features contain valuable information about the speaker identity. Hence, we create UBMs for features from sequences of two and three consecutive regions. Since the presence of pauses strongly affects the distribution of the prosodic features around them, pauses are considered as part of the sequence. Hence, feature vectors for sequences are obtained by concatenating the features or the pause length depending on whether a region is an actual syllable or pseudo-syllable or a pause. So, for example, three kinds of sequences of length 2 are defined: 1_1, 0_1 and 1_0, where a 1 indicates a syllable or pseudo-syllable and a 0 indicates a pause. For each of these sequences a separate UBM is trained.

The UBMs for each feature or group of features and each sequence are adapted to the samples independently and the obtained weights are concatenated to obtain the final transform. This transform can be shown to be a particular case of the Fisher kernel [14]. The transform is further normalized using rank normalization (described in [15]) before training the SVMs for each target speaker. For more details on this system, see [8]. No intersession variability compensation (ISVC) method is implemented for this system. Nuisance attribute projection, the most common ISVC method when training SVM models, does not lead to significant improvements on these features.

4. JFA MODELING OF GMM MEANS

In [3], Dehak et al.proposed to use JFA techniques to model the EV-PA features described in Section 2. This is possible because these features are not high-dimensional (in their paper, a total of 13 features are used) and do not contain undefined values. Using JFA to model prosodic features is a very appealing possibility, since JFA has been shown to provide outstanding performance on spectral features, and a vast amount of work has been done in the area.

JFA, as applied to speaker recognition [9], is based on the assumption that a supervector M given by the concatenated GMM means can be decomposed as $M = m + Ux + Vy + Dz$, where m is the background model supervector, U and V are low-rank matrices, D is a diagonal matrix, and x, y and z are latent variables with standard normal distribution. The components of vector x are called the *channel factors* and those of y are called the *speaker factors*. To estimate the matrix U, a database with several samples for each speaker is needed, while to estimate V a database with many different speakers is required. When the matrix D is set to 0, the model reduces to probabilistic principal component analysis (PPCA). In this paper, D is set to zero for all JFA experiments. After the parameters of the factor analysis model are computed for the train and test samples in a trial, the average log-likelihood ratio between the speaker models and the universal model are computed. Both approximate and exact ways of computing the log-likelihood have been used. In this paper we use an approximate method in which scores are computed as the scalar product between the speaker model mean offset and the channel-compensated first-order Baum-Welch (BW) statistics for the test sample centered around the UBM. This method is extremely fast and has been shown to perform very well compared to other methods [16]. The standard UBM-GMM approach corresponds to U and V set to zero and D set to correspond to maximum a posteriori (MAP) adaptation. We will call this the MAP approach.

In the case of JFA, since the model is robust to sparse data thanks to the use of speaker factors, all features for a certain polynomial degree can be modeled jointly without the need for a back-off strategy. Nevertheless, extending the work of Dehak et al., we will present results where several JFA models are trained, one for each sequence as defined in the previous section. We will show that, as in the case of the SVM modeling, sequences of length 2 and 3 give significant improvements in performance.[1] Furthermore, we will also explore the joint use of several polynomial degrees, which is, in fact, a kind of back-off strategy.

5. EXPERIMENTS

Experiments were conducted using data from the NIST SRE from 2006 and 2008, which we will call SRE06 and SRE08, respectively. SRE06 data is used for parameter tuning and combiner training, leaving SRE08 data as a clean test set. Each speaker verification trial consists of a test sample and a speaker model. The test samples are one side of a telephone conversation with approximately 2.5 minutes of speech. We consider the 1-side training conditions in which we are given one conversation side to train the speaker model. We present results on English-only and all-language subsets (subsets 6 and 7 for SRE08, as defined by NIST [17, Section 4.5]). We use ZT-NORM to normalize all scores. The data used as negative examples for the SVM training, for background model training, for ZTNORM and for training of the JFA matrices is taken from 2004 SRE data and 2005 alternate microphone SRE data. All experiments are run in a gender-dependent manner. The method used for all combination experiments is linear logistic regression where parameters are trained on SRE06 data. Results are shown in terms of equal error

[1] Marcel Kockmann from Brno University was the first to try the experiments with different sequence lengths, although no publication is yet available with his results.

rate (EER) and both the minimum and actual detection cost function (DCF), defined by NIST [17].

The size of the UBMs used for the SVM method is determined based on the number of samples available for the specific feature sequence being modeled and the dimensionality of the feature vector and goes from 24 (for single-feature models for the sequence 1_0_1) to 800 (for the joint model of all degree-5 polynomial approximation features). In the case of JFA, the number of Gaussian components used for each UBM is tuned to optimize performance on the SRE06 task matching the condition in which SRE08 results are reported (that is, for SRE08 English-only results, SRE06 data for English-only is used to optimize the number of Gaussians). Furthermore, optimization is performed separately for MAP and JFA experiments. Number of Gaussian components from 32 to 512 were explored. We find that the optimal number of Gaussians is smaller for longer sequences. Sequences of length 1 have optimal values of 256 or 512, and sequences of length 2 and 3 have optimal values between 32 and 128. The dimensions of V and U are fixed at 50 and the relevance factor for MAP at 20, since these values were found to give the best or close to best results for all models.

Figure 1 shows the minimum DCF results for four different modeling methods: MAP, JFA, SVM, and the score-level combination of SVM and JFA on SRE08 English-only data. Results are shown for each sequence length separately and for the accumulation of different sequence lengths. In the case of JFA, the systems containing sequences of length larger than 1 are obtained by score-level combination of the JFA systems for each particular sequence of pause and non-pauses of that length. Hence, for example, the system for sequence length equal to two is formed by the combination of three systems corresponding to patterns 0_0, 1_0, and 0_1. On the other hand, the SVM systems are always a single system trained with the concatenation of the features corresponding to all the patterns for the different sequence lengths involved. Furthermore, the figure presents results using a single polynomial order equal to 5, or three polynomial orders: 1, 3 and 5. As for the case of the sequence lengths, the JFA systems for different polynomial orders are obtained by score-level combination of the systems for each of the individual orders, while for the SVM system, a concatenated vector composed by the features from all orders is used to train a single SVM. In all cases, the combiner is trained on SRE06 English-only data.

We can see that the MAP results are significantly worse than the SVM results, and that JFA is significantly better than both of them for all conditions. Furthermore, the combination of the JFA and SVM systems leads to further improvements. In all cases we see that adding sequences of higher lengths leads to significant performance improvements. Sequences of length 3 alone are worse than those of length 2. This is likely to be due to the increased dimensionality of the feature vectors, which makes it harder to robustly estimate the UBMs. Interestingly, the addition of lower polynomial orders also leads to improvements, indicating that even JFA techniques can benefit from back-off strategies.

The results highlighted with stars correspond to the two results previously presented in the literature. The red star roughly corresponds to the results presented in [3], while the green star roughly corresponds to the results presented in [8]. We can see that, once we extend the JFA method by including some of the characteristics of the original SVM system, the DCF improves from 0.66 to 0.50, a 25% relative improvement. Furthermore, the combination of both methods outperforms the best of the two previous baselines (green star) by 30%.

There might be circumstances in which JFA cannot be used due to lack of appropriate data with which to train the matrices for the model. Our results indicate that the SVM method might be preferable over the MAP method for such a situation, since the SVM method seems to be inherently more robust to session variability than the UBM-GMM method when using MAP to estimate the speaker's model. Further experiments should be run to confirm this conclusion, though, including exact scoring methods, careful tuning of the relevance factor to the particular task, etc.

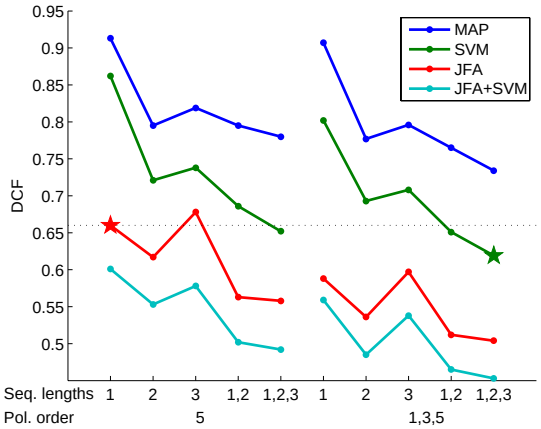

Fig. 1. Results for four different modeling methods: MAP, JFA, SVM, and JFA combined with SVM on SRE08 English-only data. Two sets of results are shown, using only polynomial order 5 and using 1, 3 and 5. For each of them, results are shown for each sequence length separately, and for the accumulation of different sequence lengths. The two stars correspond to the two results previously published in the literature.

Table 1 shows a summary of results, for several prosodic systems and their combination with a baseline system. The baseline system used here is a JFA system based on 20 MFCC features plus deltas and double deltas. The data used to train the UBM and the ZTNORM lists for this system come from 2004 and 2005 alternate microphone SRE data. For JFA, that data plus Switchboard 2 data Phase 2, 3 and 4, and the development interview data released for SRE08 are used. Results for our best prosodic system to date, which we call the *full prosodic system*, are also shown in the table. This system uses all SNERF features, grammatically constrained SNERFs, and the EV-PA features modeled together in a single SVM system. The system uses the English subset of the data used for UBM training and ZTNORM for the other prosodic systems in this paper. This system is described in detail in [2]. Since this system requires the output of an ASR system, only English results are available for it.

The table shows that even after the large gains in the prosodic system presented in this paper, the full prosodic system which uses an order of magnitude more features and includes features with undefined values that cannot be handled under the current JFA framework, is still significantly better. Overall, improvements of around 10% on both EER and DCF can be achieved on the English conditions when combining the baseline system with a prosodic system. Furthermore, we see that gains in the performance of the prosodic system correspond to gains (though, as usual, much more modest) in the combination performance. Notably, while the original SVM prosodic system gives no improvement in combination with the baseline for the all-language conditions, a gain is observed when the improved prosodic system is used. We believe larger gains can be achieved if more data (comparable to that used for the baseline system) is used for UBM and JFA training, and for the ZTNORM lists for the prosodic systems.

System	SRE06				SRE08					
	Eng (23687)		All lang (51068)		Eng (17761)			All lang (35896)		
	mDCF	EER	mDCF	EER	aDCF	mDCF	EER	aDCF	mDCF	EER
SVM O:1,3,5 N:1,2,3	0.603	14.518	0.670	17.533	0.632	0.619	15.961	0.854	0.812	19.753
JFA O:5 N:1	0.649	12.188	0.687	14.242	0.669	0.660	14.577	0.812	0.773	16.990
JFA O:1,3,5 N:1,2,3	0.468	9.372	0.541	11.753	0.528	0.504	11.645	**0.673**	0.644	14.339
JFA + SVM O:1,3,5 N:1,2,3	0.425	8.451	**0.508**	**11.394**	0.467	0.453	10.831	0.687	**0.634**	**13.667**
Full prosodic system	**0.336**	**7.313**	-	-	**0.417**	**0.407**	**9.528**	-	-	-
Baseline	**0.081**	**1.679**	**0.142**	**2.987**	**0.116**	**0.117**	**2.687**	**0.331**	**0.331**	**5.788**
Base + SVM O:1,3,5 N:1,2,3	0.076	1.734	0.143	2.959	0.116	0.113	2.687	0.378	0.331	5.825
Base + JFA O:5 N:1	0.077	1.679	0.139	2.849	0.115	0.114	2.443	0.358	0.323	5.601
Base + JFA O:1,3,5 N:1,2,3	0.076	1.679	0.137	2.821	0.115	0.111	2.443	0.355	0.319	5.489
Base + JFA + SVM O:1,3,5 N:1,2,3	0.075	1.734	**0.136**	**2.765**	0.115	0.110	2.443	**0.351**	**0.311**	**5.452**
Base + Full prosodic system	**0.069**	**1.517**	-	-	**0.100**	**0.097**	**2.362**	-	-	-

Table 1. Results on SRE08 for different prosodic systems alone and in combination with a baseline spectral system. The order of the EV-PA features (O) and the sequence lengths (N) used are indicated for each system. Minimum DCF (mDCF) and equal error rate (EER) are shown for SRE06 data. Actual DCF (aDCF) with threshold estimated on SRE06 data is also shown for SRE08 data. The number in parenthesis beside the language condition indicates the number of trials.

6. CONCLUSIONS

We presented a study of two different modeling methods, JFA modeling of GMM means and SVM modeling of GMM weights, for a subset of simple prosodic features obtained by polynomial approximations of the pitch and energy signals over pseudo-syllables. Our results indicate that, for these features, the JFA method greatly outperforms the SVM method, and that the combination of both methods leads to significant gains over JFA alone. Our results extend the previous use of JFA for these features, including the modeling of different sequence lengths and different polynomial order approximations. We demonstrate a gain of 25% on the JFA method after these additions.

Even though the JFA method clearly outperforms the SVM method for the simple features used here, it is not clear whether this method can be used for the more general set of prosodic features, which are still shown to outperform the best result obtained with the simple set of features. The adaptation of JFA techniques to the larger feature set is thus an important area for future research.

7. ACKNOWLEDGMENTS

This research was funded by NSF CNS-0652510 at Stanford University and through a development contract with Sandia National Laboratories by the National Geospatial-Intelligence Agency (NGA) under National Technology Alliance (NTA) Agreement Number NMA 401-02-9-2001 at SRI International. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NSF, NGA, the United States Government, or Rosettex.

8. REFERENCES

[1] L. Ferrer, E. Shriberg, S. Kajarekar, A. Stolcke, K. Sönmez, A. Venkataraman, and H. Bratt, "The contribution of cepstral and stylistic features to SRI's 2005 NIST speaker recognition evaluation system," in *Proc. ICASSP*, Toulouse, May 2006, vol. 1, pp. 101–104.

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

[3] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, Sept. 2007.

[4] Marcel Kockmann and Lukas Burget, "Contour modeling of prosodic and acoustic features for speaker recognition," in *Proceedings of 2008 IEEE Workshop on Spoken Language Technology*, Goa, IN, Dec. 2008.

[5] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICSLP*, Sydney, Dec. 1998, vol. 7, pp. 3189–3192, Australian Speech Science and Technology Association.

[6] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02," in *Proc. ICASSP*, Hong Kong, Apr. 2003, vol. 4, pp. 792–795.

[7] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46, no. 3-4, pp. 455–472, 2005, Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation.

[8] L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sönmez, "Parameterization of prosodic feature distributions for SVM modeling in speaker recognition," in *Proc. ICASSP*, Honolulu, Apr. 2007.

[9] 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.

[10] Chi-Yueh Lin and Hsiao-Chuan Wang, "Language identification using pitch contour information," in *Proc. ICASSP*, Philadelphia, Mar. 2005, vol. 1, pp. 601– 604.

[11] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. ICSLP*, Beijing, Oct. 2000, China Military Friendship Publish.

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

[13] S. Kajarekar, L. Ferrer, K. Sönmez, J. Zheng, E. Shriberg, and A. Stolcke, "Modeling NERFs for speaker recognition," in *Proc. Odyssey-04*, Toledo, Spain, May 2004, pp. 51–56.

[14] L. Ferrer, *Statistical Modeling of Heterogeneous Features for Speech Processing Tasks*, Ph.D. thesis, Stanford University, 2009.

[15] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification," in *Proc. ICASSP*, Las Vegas, Apr. 2008.

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

[17] "NIST SRE08 evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf.