RECENT PROGRESS IN PROSODIC SPEAKER VERIFICATION

Marcel Kockmann[1], Luciana Ferrer[2], Lukáš Burget[1], Elizabeth Shriberg[2] and Jan "Honza" Černocký[1]

[1]Brno University of Technology, Speech@FIT, Czech Republic
[2]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA

ABSTRACT

We describe recent progress in the field of prosodic modeling for speaker verification. In a previous paper, we proposed a technique for modeling syllable-based prosodic features that uses a multinomial subspace model for feature extraction and within-class covariance normalization or linear discriminant analysis for session variability compensation. In this paper, we show that performance can be significantly improved with the use of probabilistic linear discriminant analysis (PLDA) for session variability compensation. This system does not require score normalization. We report an equal error rate below 7% on a NIST 2008 task. To our knowledge, this is the best reported result to date for a prosodic system for speaker recognition. Fusion of this system with a state-of-the-art acoustic baseline system yields 10% relative improvement in the new detection cost function (DCF) as defined by NIST.

Index Terms— Prosodic speaker verification, SNERFs, MSM, iVector, PLDA

1 INTRODUCTION

Using high-level information to further enhance short-time, cepstral-based speaker verification systems has been popular for several years. In [1], several high-level features (phonetic, prosodic, linguistic, etc.) were leveraged to enhance the Equal Error Rate (EER) on the NIST 2001 speaker recognition evaluation task up to 70% relative. This gain from using high-level features was enabled by the introduction of evaluation conditions with large train and test durations (of 2.5 minutes for testing and up to 8 times that amount for training). High-level features are sparser than lower-level acoustic features and, hence, benefit more from large amounts of data. During subsequent NIST evaluations, challenging new corpora and rapid performance improvements for systems using standard cepstral features generally made gaining an advantage from the fusion of high-level features difficult [2].

Nevertheless, in 2004, high-level features were shown to provide performance gains greater than 30% when combined with a baseline acoustic system on the NIST 2004 tasks [3]. The success was mainly due to SRI's newly proposed, syllable-based, non-uniform extraction region features (SNERFs) [4]. These features in combination with specialized parameterization methods and Support Vector Machine (SVM) modeling [5] resulted in the best-performing prosodic system at the time. But, the SNERF system was complex and for this reason, was not broadly adopted by the community.

The introduction of joint factor analysis (JFA) [6] for speaker verification brought the performance of acoustic systems for speaker recognition to a new level, leading to improvements on the order of 50% over previous state-of-the-art systems. As a consequence of these dramatic improvements in the baseline performance of speaker recognition systems, obtaining gains from high-level features, particularly if they could not capitalize on the JFA improvements obtained for acoustic systems, was increasingly difficult. A first step in using JFA for prosodic systems was proposed by [7] for a set of very simple prosodic features. This framework for prosodic modeling has been adopted by several sites and investigated thoroughly [8, 9].

Unfortunately, the JFA framework cannot be directly applied to the SNERFs due to their high dimensionality and to the existence of undefined values. In [9], we showed that the SNERF system still outperforms a simpler set of features modeled with JFA. This was our motivation for trying to transfer the underlying idea of JFA – to model speaker and intersession variability in low-dimensional subspaces – to a model that can handle SNERFs. Recently, we presented a theoretic framework for the modeling of SNERFs using a multinomial subspace model (MSM), which achieved very promising results [10].

This paper describes our latest progress in using Probabilistic Linear Discriminant Analysis (PLDA) modeling for session variability compensation of features obtained with MSM. Significant gains are achieved over previous performance, resulting in an equal error rate (EER) of 6.9% on the telephone data of the NIST 2008 Speaker Recognition Evaluation [11]. To our knowledge, these are the best results in the literature for a prosodic speaker verification system. Furthermore, no score normalization techniques are needed. In addition, we present fusion experiments with a state-of-the-art acoustic JFA system showing gains of up to 10% in detection cost function (DCF). A major goal of this paper is to clearly describe the complex system-building process. All important steps – from raw SNERF features to final PLDA modeling – are explained in Section 2. In Section 3, our experimental setup is described and different prosodic systems are evaluated and compared. Fusion results with a baseline acoustic system are also shown. We present our conclusions in Section 4.

2 SYSTEM

This section describes the five major steps of the system-building process. All steps are explained using a simplified example. Please refer to the citations for algorithmic descriptions.

2.1 Syllable-based NERFs (SNERFs)

We use SNERFs [4], which are syllable-based, non-uniform extraction region features based on F0, energy, and duration information. Characteristics like minimum, maximum, mean, and slope of the

This work was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U. S. Government. The work was also partly supported by European project MOBIO (FP7-214324), Grant Agency of Czech Republic project No. 102/08/0707, Czech Ministry of Education project No. MSM0021630528 and by BUT FIT grant No. FIT-10-S-2. Marcel Kockmann was supported by SVOX Deutschland GmbH.

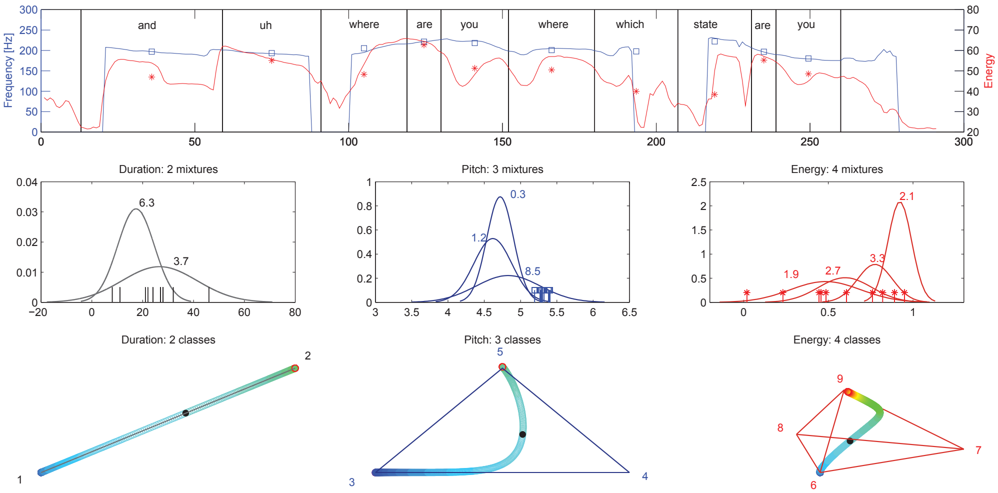

Fig. 1. **Top row:** Extraction of three SNERF parameters from a speech segment containing 10 single-syllable words: Syllable duration (determined by black vertical lines), mean pitch value per syllable (blue squares), and mean energy per syllable (red stars). **Middle row:** Parameterization of SNERF sequences: Small GMMs are trained on background data for each individual SNERF. Two mixtures are used for duration, three mixtures for pitch, and four mixtures for energy. Occupation counts for the values extracted in the top row (here as bars) are collected using the GMMs. **Bottom row:** Multinomial model spaces for duration, pitch, and energy. The colored lines show various one-dimensional iVectors (the values are mapped to colors) projected to the full ensemble of multinomial spaces.

pitch and energy trajectories are extracted for each detected syllable in an utterance and for its nucleus, as well as the duration of onset, nucleus and coda of the syllable. All values are further normalized with different techniques, resulting in a few hundred features for each syllable (174 in our current implementation). The syllable segmentation is generated from the output of a large vocabulary continuous speech recognition (LVCSR) system. The phone alignments of the recognized words are used to generate English syllables. Detailed information on SNERFs is given in [4].

Temporal dependencies are modeled by concatenating features from consecutive syllables and pauses. New vectors are formed for each basic feature by concatenating consecutive values. If a pause is found within the sequence, the length of the pause is used as a feature. For each sequence length, each feature, and each pattern of pause/non-pause, we obtain a separate feature vector. For example, for trigrams, we obtain five different vectors: (S, S, S), (P, S, S), (S, P, S), (S, S, P), (P, S, P) for each feature. Each pair $\{$feature, pattern$\}$ determines what we call a *token* (see [5] for details). Our current implementation uses sequences of lengths 1, 2, and 3. The first line of plots in Figure 1 shows an example of the feature extraction process. The segments are given by the syllables found from the ASR output. The pitch (blue curve) and energy (red curve) signals are estimated from the waveform. For our example, we assume that we extract only three features per segment: its duration (from one vertical black line to the next), the mean pitch value (blue squares), and the mean energy value (red stars).

2.2 Background GMMs

For each token, we train a separate Gaussian Mixture Model (GMM) with a small number of mixture components on the background data. Because basic features may be undefined (e.g., when no pitch is detected or when the syllable lacks onset or coda), a special GMM is needed using an additional parameter for the probability of a feature being undefined. In the first pass, all GMMs are trained using frames with defined features only, where the additional parameter is set to one and the model falls back to a standard GMM. The GMMs are then retrained with all feature vectors, allowing the new parameter to adapt to the data. Details of the modified expectation-maximization algorithm are given in [12]. The second line of Figure 1 shows a toy example in which three small GMMs are trained on a background data set. A two-component model is trained for the syllable durations, a three-component model for mean pitch values, and a four-component GMM for means of syllable energies.

2.3 Parameterization of SNERF sequences

After training the background models for each token, we gather Gaussian component occupation counts for each utterance (zero order sufficient statistics from the modified EM algorithm [12]). These are accumulated soft counts describing the responsibilities of each individual mixture component toward generating the frames in the utterance. Using these parameters, we transform the sequence of SNERFs (one feature vector per syllable) to fixed length vectors (one vector of statistics per utterance). The values from the exemplified feature extraction process (syllable duration, mean pitch, and mean energy) are further depicted as bars in the middle row of Figure 1. The occupation counts (the numbers next to the mixtures) are the responsibilities for each Gaussian component in generating these values. Each Gaussian component can be seen as a discrete class and the occupation counts can be seen as soft-counts of discrete events.

2.4 Multinomial Subspace Model

As a generative model, a multinomial distribution appears as a natural choice for modeling the counts resulting from the previous step. More precisely, a set of E multinomial distributions is required, one for each GMM in the ensemble. Each multinomial distribution corresponds to a set of C_e probabilities, one probability ϕ_{ec} for each Gaussian c in the GMM e. For each frame, each GMM is expected to generate a feature by one of its components with probability given by the multinomial distribution. This corresponds to co-occurring events that should be modeled by separate multinomial distributions (as all tokens are modeled independently of each other). Each multinomial distribution lives in a n-dimensional simplex and the space of all parameters is the cartesian product of all the simplexes. The bottom row of Figure 1 illustrates this for our toy example where the parameters of the duration model exist on a line; the pitch model parameters, in a 2D simplex; and the energy parameters, in a 3D simplex space.

We use a Multinomial Subspace Model (MSM) [10] where we assume that the multinomial distributions differ from utterance to utterance. In the case of SNERFs, we need to estimate parameters of many multinomial distributions. Therefore, we search for a way to estimate all the parameters robustly given a limited amount of data available for each utterance. With MSM, we assume that there is a low-dimensional subspace of the parameter space in which the parameters for individual utterances live. For this reason we introduce an explicit latent variable $\mathbf{w}$ through which the probability ϕ_{ec} of cth class of each multinomial distribution e in the ensemble is given by

$$\phi_{ec} = \frac{\exp(\mathbf{t}_{ec}\mathbf{w})}{\sum_{i=1}^{C_e} \exp(\mathbf{t}_{ei}\mathbf{w})}, \tag{1}$$

with $\mathbf{t}_{ec}$ being the cth row of eth block of subspace matrix $\mathbf{T}$ (size $\sum_{e=1}^{E} C_e \times r$) which spans a linear subspace that might be nonlinear in the original parameter space due to the softmax function. Figure 1 shows how the subspace restricts the movement in the full parameter-space in a non-linear way (colored lines). By drawing values for a one-dimensional variable $\mathbf{w}$ from minus infinity to infinity we move in all three simplexes simultanuously along the non-linear, low-dimensional manifolds. Now, all the multinomial distributions corresponding to one utterance can be represented by a low dimensional vector $\mathbf{w}$. This way, we can (1) reduce the number of free parameters to efficiently model differences between individual utterances, and (2) learn dependencies between the individual SNERFs.

The MSM parameters are estimated by iteratively re-estimating the latent variables $\mathbf{w}$ for each utterance in the training data to maximize the likelihood function based on the current estimate of $\mathbf{T}$ and vice-versa. Using the final estimate of $\mathbf{T}$ we can extract $\mathbf{w}$ vectors (which we will call iVectors) for new data. This way, the MSM is used as a feature extractor and each iVector can be seen as a low-dimensional representation of the whole utterance.

2.5 PLDA modeling

For verification of speaker trials we use a special case of Probabilistic Linear Discriminant Analysis (PLDA) [13], a two-covariance model, providing a probabilistic framework where speaker and inter-session variability in the iVectors is modeled using across-class and within-class covariance matrices $\mathbf{\Sigma}_{ac}$ and $\mathbf{\Sigma}_{wc}$. We assume that latent vectors $\mathbf{y}$ representing speakers are distributed according to

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mu, \mathbf{\Sigma_{ac}}) \tag{2}$$

and for a given speaker $\mathbf{y}$ the iVectors are distributed as

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}; \mathbf{y}, \mathbf{\Sigma_{wc}}). \tag{3}$$

Model parameters μ, $\mathbf{\Sigma}_{ac}$ and $\mathbf{\Sigma}_{wc}$ are trained using an EM algorithm [14]. Using the PLDA model, one can directly evaluate the log-likelihood ratio for the hypothesis test corresponding to “the two iVectors were generated by the same speaker or not”:

$$s = \log \frac{\int p(\mathbf{w}_1|\mathbf{y})p(\mathbf{w}_2|\mathbf{y})p(\mathbf{y})d\mathbf{y}}{p(\mathbf{w}_1)p(\mathbf{w}_2)} \tag{4}$$

The numerator gives the marginal likelihood of producing both iVectors from the same speaker, while the denominator is the product of the marginal likelihoods that both iVectors are produced from different speakers. The integrals can be evaluated analytically and scoring can be performed very efficiently as described in [15].

3 EXPERIMENTS AND RESULTS

This section describes our results for three individual prosodic systems, including two previously-proposed systems. We also show results when fusing the prosodic systems with a state-of-the-art cepstral system.

3.1 Data

The task used to present results uses data from the NIST 2008 speaker recognition evaluation. The original NIST tasks are extended to include two orders of magnitude more impostor samples. This was done to support the new DCF metric introduced by NIST for the 2010 evaluation [16]. In this paper, we show results only for the telephone condition, in which both training and test samples are given by telephone conversations recorded over a telephone channel. The number of target and impostor samples for this task are 1,108 and 1,453,237, respectively. As background data to train UBMs, JFA, MSM and PLDA we use data from the 2004 and 2005 SRE, 2008 interview development data and from the Switchboard-II corpus.

3.2 Prosodic systems

We evaluate three different prosodic systems: (1) a system based on JFA-modeling of means of low-dimensional polynomial features (Prospol) describing pitch and energy trajectories, originally proposed in [7] and further extended and improved as described in [9]; (2) the baseline SNERF system with SVM modeling (SNERF-SVM) of the counts as originally described in [5]; and (3) the recently introduced subspace model [10] with additional PLDA modeling applied to the SNERF counts (SNERF-IV-PLDA).

The Prospol system models a small set of 13 features, including polynomial approximations of the pitch and energy profiles and the duration of the region for three different region definitions: (1) energy valleys (as originally proposed in [7]); (2) uniform windows of 300 msec shifted by 10 msec (as proposed in [8]); and (3) syllable regions (identical to those used for the SNERFs). Further, sequences of length 2 are also modeled. For each region and each sequence length, a separate system is created. The resulting scores are combined with fixed weights determined empirically from development data. The baseline SNERF system directly uses the occupation counts (divided by the number of frames) as features for an SVM model (steps 2.1–2.3). Session variability compensation can be applied to this model using nuisance attribute projection [17], but we found no significant gains from this approach. For the SNERF-IV-PLDA system the occupation counts are used to train an MSM with a subspace dimension r=200 following [10]. Next, iVectors are extracted using this model for all background, training and test utterances. The PLDA model is then trained[1] on iVectors extracted for all background data and is used to perform verification between speaker trials. Figure 2 shows the DET curves for the three prosodic systems. Both SNERF systems outperform the Prospol system at all operating points of interest. Further, the proposed modeling technique for the SNERFs is significantly better than the older method based on SVMs for most operating points resulting in an EER of 6.9%. Moreover, the PLDA modeling significantly outperforms the cosine distance scoring with LDA as used in our previous work with MSMs (9% EER) [10].

3.3 Acoustic system

The cepstral GMM baseline system uses a 300-3300 Hz bandwidth front-end consisting of 24 Mel filters to compute 20 cepstral coeffi-

[1] We thank Niko Brümmer for providing his PLDA implementation.

Table 1. *Relative improvement over cepstral JFA baseline [%].*

	System	new DCF	old DCF	EER
Fusion	Baseline+Prospol	6.25	-1.37	-5.26
	Baseline+SNERF-SVM	7.21	3.70	10.53
	Baseline+SNERF-IV-PLDA	9.62	5.08	5.27

cients with cepstral mean subtraction, and their delta, double delta coefficients, producing a 60-dimensional feature vector. The resulting features are mean- and variance-normalized over the utterance. The feature vectors are modeled by a 1024-component, gender-independent GMM. We use a full Joint Factor Analysis model (JFA) in which 600 eigenvoices are trained and 250 eigenchannels are trained separately for telephone and interview data and are concatenated. The diagonal term is trained with the same data as used to train the speaker factors. Scores are normalized using gender-dependent ZTnorm, resulting in an EER of 1.65%, an old DCF of 0.073, and a new DCF of 0.42.

3.4 Fusion

Fusion results are obtained using a cross-validation paradigm. To this end, the complete set of speakers is split into two disjoint sets. The trials involving only speakers from each of these sets are then selected. In the process, half of the impostor trials (those corresponding to one speaker from one set and another speaker from the other set) are discarded. The fusion parameters are then trained using standard linear logistic regression on one of the sets and then applied to the other set, and conversely. The results shown in Table 1 are computed on the concatenation of these two sets. The fusion results show that the SNERF systems result in larger and more consistent gains over the baseline. This justifies using the SNERF features over the simpler polynomial features. Further, even though both SNERF systems give somewhat similar gains in combination, the proposed modeling technique should be more robust to noisy conditions and other types of variabilities, because the SNERF-SVM approach does not implement any kind of session variability compensation.

4 CONCLUSION

We have proposed a technique for modeling complex prosodic features, such as SNERFs, using a multinomial subspace model for feature extraction and probabilistic linear discriminant analysis for session variability compensation. The proposed system achieves more than 20% relative improvement with respect to the current prosodic systems on EER and old DCF metrics. An interesting finding is that the large gains from the proposed modeling technique decrease as the cost metric moves toward the low false acceptance region. In fact, at the recently introduced new DCF metric, which corresponds to very low false acceptance rates, both SNERF systems perform similarly. Comparing the performance of the polynomial prosodic features to the SNERFs, we see that SNERFs greatly outperform the simpler features. This behavior requires further investigation to understand whether it is due to the difference in the nature of the features, to the new modeling technique, or to both factors. Although SNERFs cannot be modeled with JFA, polynomial features could be modeled using the proposed MSM/PLDA technique. However, initial results in this direction did not show gains with respect to JFA modeling for these features.

In the future, we plan to investigate the performance of prosodic systems on diverse channel conditions and for different speech styles (interview conversations and telephone calls recorded over microphones other than telephone handsets). Further investigation is also needed to understand the influence of the subspace size. Finally, we plan to explore the use of heavy-tailed distributions in PLDA [14], which has been shown to give significant improvements for acoustic systems.

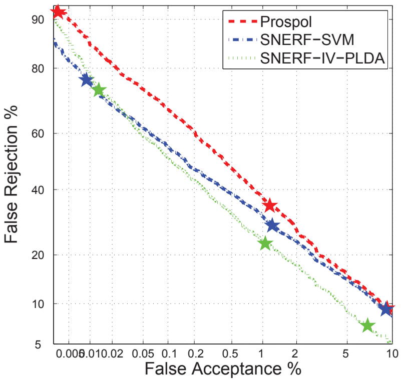

Fig. 2. DET curves for the three prosodic systems. The three markers in each line correspond to the new DCF, the old DCF, and the EER (as used by NIST to evaluate SRE 2008 [11] and 2010 [16]), from left to right.

5 References

[1] D. Reynolds *et al.*, “The SuperSID project: Exploiting high-level information for high-accuracy,” in *in Proc. International Conference on Audio, Speech, and Signal Processing, Hong Kong*, 2003, pp. 784–787.

[2] ——, “The 2004 MIT lincoln laboratory speaker recognition system,” pp. 177 – 180, 2005.

[3] S. S. Kajarekar *et al.*, “SRIs 2004 nist speaker recognition evaluation system,” in *in Proc. ICASSP*, 2005, pp. 173–176.

[4] E. Shriberg *et al.*, “Modeling prosodic feature sequences for speaker recognition,” *Speech Communication*, Jan 2005.

[5] L. Ferrer *et al.*, “Parameterization of prosodic feature distributions for SVM modeling in speaker recognition,” *Proc. ICASSP, Taipei*, vol. 4, pp. 233–236, 2007.

[6] P. Kenny *et al.*, “A study of inter-speaker variability in speaker verification,” *IEEE Trans. Audio*, Jan 2008.

[7] N. Dehak *et al.*, “Modeling prosodic features with joint factor analysis for speaker verification,” *Audio, Speech and Language Processing*, Jan 2007.

[8] M. Kockmann *et al.*, “Investigations into prosodic syllable contour features for speaker recognition,” *Proc. of ICASSP, Dallas*, Sep 2010.

[9] L. Ferrer *et al.*, “A comparison of approaches for modeling prosodic features in speaker recognition,” in *Proc. ICASSP, Dallas*, 2010.

[10] M. Kockmann *et al.*, “Prosodic speaker verification using subspace multinomial models with intersession compensation,” in *Proc. Interspeech, Tokyo*, 2010.

[11] NIST, “The NIST year 2008 speaker recognition evaluation plan,” pp. 1–10, Apr 2008.

[12] S. Kajarekar *et al.*, “Modeling NERFs for speaker recognition,” in *Proc. Odyssey, Toledo*, 2004, pp. 51–56.

[13] S. J. D. Prince, “Probabilistic linear discriminant analysis for inferences about identity,” in *ICCV*, 2007.

[14] P. Kenny, “Bayesian speaker verification with heavy tailed priors,” in *Keynote presentation, Odyssey*, 2010.

[15] L. Burget *et al.*, “Discriminatively trained probabilistic linear discriminant analysis for speaker verification,” in *ICASSP*, 2011.

[16] NIST, “The NIST year 2010 speaker recognition evaluation plan,” 2010. [Online]. Available: http://www.itl.nist.gov/iad/mig//tests/sre/2010

[17] A. Solomonoff *et al.*, “Channel compensation for SVM speaker recognition,” in *Odyssey*, 2004, pp. 57–62.