iVector Fusion of Prosodic and Cepstral Features for Speaker Verification

Marcel Kockmann[1] and Luciana Ferrer[2] and Lukáš Burget[1] 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

In this paper we apply the promising iVector extraction technique followed by PLDA modeling to simple prosodic contour features. With this procedure we achieve results comparable to a system that models much more complex prosodic features using our recently proposed SMM-based iVector modeling technique. We then propose a combination of both prosodic iVectors by joint PLDA modeling that leads to significant improvements over individual systems with an EER of 5.4% on NIST SRE 2008 telephone data. Finally, we can combine these two prosodic iVector front ends with a baseline cepstral iVector system to achieve up to 21% relative reduction in new DCF.

Index Terms: speaker verification, prosody, JFA, iVector, SMM, fusion

1. Introduction

High-level information has been used for over a decade to further enhance short-time, cepstral-based speaker verification systems. Many approaches make use of acoustic attributes of speech prosody that mainly involve variations in syllable length, loudness, and pitch. In recent NIST Speaker Recognition Evaluations [1, 2], two families of prosodic feature sets were presented. One family corresponds to syllable-based, non-uniform extraction region features (SNERFs) [3], which are highly complex prosodic features originally proposed by SRI. These features in combination with specialized parameterization methods and support vector machine (SVM) modeling [4] result in a very good prosodic system.

Another family of systems uses a set of very simple prosodic features, originally proposed for language identification [5]. These features model the temporal trajectory of pitch and energy over the time span of a syllable. Joint Factor Analysis (JFA) modeling for these features was originally proposed by [6] and showed very promising results. This framework for prosodic modeling has been adopted by several sites and investigated thoroughly [7, 8]. The main reason for its success lies in JFA modeling, which is capable of coping with the problem of speaker and session variability in Gaussian mixture model (GMM)-based speaker verification [9] and has become the de facto standard for modeling low- and high-level features.

Moreover, excellent results on cepstral features were obtained with a simplified variant of JFA [10], where separate subspaces for channel and speaker variability are replaced by a single subspace covering the total variability. This model can be used to extract compact low-dimensional feature vectors representing a whole utterance, often called iVectors. Based on this idea, we proposed a framework where the subspace modeling technique normally used to model means of GMMs is adapted to model occupation counts using a multinomial model. This so-called Subspace Multinomial Model (SMM) [11] is applicable to the complex SNERFs to extract iVectors.

Probabilistic Linear Discriminant Analysis (PLDA) [12] has been proposed to model the speaker and channel variability in both types of iVectors, directly generating likelihood ratios for the trials [13, 14]. iVector modeling of SNERFs by SMMs with successive PLDA has been shown to give the best results for a prosodic speaker verification system so far [15].

To date, the iVector approach – using a total variability subspace followed by PLDA – has not been used (to our knowledge) for the simple prosodic features that are usually modeled by JFA.

In this paper, we present results on the prosodic JFA system as presented by Brno University of Technology in SRE 2010 and apply iVector modeling and PLDA back end to the same features. We show that the iVector approach is superior to the standard JFA modeling even for simple prosodic features.

In this way we have two diverse prosodic systems that achieve similar performance on our test sets: an iVector system that models means of GMMs based on simple well-defined prosodic features and an iVector system that models counts of multinomial distributions based on SNERFs. A combination of both systems seems relevant due to their complementary nature in terms of features and modeling. We propose an elegant way of combining these systems by simple concatenation of individual iVectors followed by a single joint PLDA model. This combination achieves an equal error rate (EER) of 5.4% on our NIST SRE 2008 telephone test set, a 23% gain over the best of the two systems.

Justification for use of a higher-level systems usually lies in an overall improvement by fusion with a cepstral baseline system. Usually, combination of low- and high-level systems is done by score-level fusion using a separate development set to train the fusion parameters. As the best-performing cepstral systems to date are also based on iVector modeling followed by PLDA modeling [13, 14, 16], we are inspired by the successful combination of two prosodic iVector front ends to further combine the cepstral and prosodic systems in the same manner. We achieve a relative reduction in terms of the challenging new detection cost function (DCF) [2] of 17% for SRE 2010 data and 21% for SRE 2008 data. The iVector combination consistently outperforms standard score-level fusion (11% and 13%) with no need for a separate development set to train the fusion parameters.

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 Czech Ministry of Education project No. MSM0021630528, and Grant Agency of Czech Republic project No. GP102/09/P635. Marcel Kockmann was supported by SVOX Deutschland GmbH.

2. Prosodic features

This section describes the two prosodic feature sets used in the paper.

2.1. DCT contour features

The DCT contour feature generation closely follows the description in [7]. The features incorporate duration, pitch and energy measurements. Pitch and energy values are estimated every 10 ms, and energy is further normalized by its maximum. The temporal trajectory of pitch and energy is modeled by a discrete cosine transform (DCT), over a fixed frame long temporal window of 300 ms, with a 50 ms frame shift. The first six DCT coefficients of both pitch and energy trajectories form a fixed-length feature vector. Only voiced frames (where pitch is detected) are used to estimate the DCT. Duration information measured as the number of voiced frames within the 30-frame interval is appended and treated as a continuous value when modeling the distributions.

2.2. SNERF features

We use SNERFs, which are syllable-based prosodic features based on estimated pitch, energy, and duration information. Characteristics like minimum, maximum, mean, and slope of pitch and energy trajectories are extracted for each detected syllable in an utterance and its nucleus, as well as duration of onset, nucleus, and coda of the syllable. All values are further normalized with different techniques and form several hundred features for each syllable. The used syllable segmentation is generated from the output of a large-vocabulary continuous speech recognition (LVCSR) system using a simple maximum onset algorithm (Section 3.4.1 of [17]) on the phone-level alignments. Detailed information on SNERFs is given in [3].

We use 182 basic features that are extracted for each syllable. Furthermore, temporal dependencies are modeled by constructing small vectors concatenating features from consecutive syllables and pauses. These so-called tokens are formed for each basic feature by concatenating as many as three values (feature values and duration of pauses; more details are given in [4]). Nine different n-gram tokens are used.

The SNERFs are parameterized by use of GMMs. This can be seen as a soft binning of each SNERF value into a meaningful set of discrete classes and makes it possible to accumulate soft counts for all SNERFs and tokens extracted for one utterance (for details see [4]).

3. Subspace models for prosodic features

The basic assumption in subspace modeling is that the natural parameters of a model usually live in a much smaller subspace than the full parameter space. This subspace can be learned by introducing latent variables in the model.

3.1. iVectors based on GMMs

The classical formulation of JFA for speaker verification [9] assumes that the concatenated mean vectors ϕ_{GaussJFA} of a GMM are distributed according to a subspace model with separate subspaces for speaker and channel variability:

$$\boldsymbol{\phi}_{\text{GaussJFA}} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x}, \tag{1}$$

where $\mathbf{m}$ is a speaker- and channel-independent supervector, and $\mathbf{V}$ and $\mathbf{U}$ span linear subspaces (for speaker and channel variability) in the original mean parameter space. The components of $\mathbf{y}$ and $\mathbf{x}$ are the low-dimensional latent variables corresponding to the speaker and channel subspaces.

A simplified variant of JFA [10] assumes that speaker and channel subspaces are not decoupled and uses only one subspace covering the total variability in an utterance:

$$\boldsymbol{\phi}_{\text{GaussIV}} = \mathbf{m} + \mathbf{T}\mathbf{w}. \tag{2}$$

Again, $\mathbf{T}$ spans a linear subspace in the original mean parameter space and the components of $\mathbf{w}$ are the low-dimensional latent variables corresponding to the total variability subspace. The low-dimensional vectors $\mathbf{w}$ are also known as iVectors.

In the latter approach, the JFA-like model serves only as the extractor of the vectors $\mathbf{w}$, which can be seen as low-dimensional fixed-size representations of utterances, and which are in turn used as inputs to another classifier.

Both techniques, the JFA (*GaussJFA*) as well as the iVector modeling (*GaussIV*), are applicable to mean supervectors of GMMs trained on the low-dimensional well-defined DCT features as presented in Section 2.1. All model parameters are trained using an expectation-maximization (EM) algorithm [9].

3.2. iVectors based on multinomial distributions

The weights of a GMM can also be modeled under the subspace paradigm. To do this, we consider the individual mixture components in the GMM to be discrete classes which can be modeled using a multinomial distribution. Similar to *GaussIV*, SMM assumes that there is a low-dimensional subspace of the parameter space in which the parameters of the multinomial distributions for individual utterances live. The probability ϕ_{MultinIV} of cth class of the multinomial distribution is given by

$$\phi_{\text{MultinIV}} = \frac{\exp(m + \mathbf{t}_c\mathbf{w})}{\sum_{i=1}^{C} \exp(m + \mathbf{t}_i\mathbf{w})}, \tag{3}$$

where $\mathbf{w}$ is a latent variable and $\mathbf{t}_c$ is the cth row of subspace matrix $\mathbf{T}$, which spans a linear subspace in the log-probability domain. Due to the softmax function, this corresponds to a possibly nonlinear subspace in the simplex that the multinomial distributions live in.

Given the parameters $\mathbf{m}$ and $\mathbf{T}$ we can extract $\mathbf{w}$ vectors (which we will also call iVectors) for new data. Similar to the *GaussIV* system, the SMM is used as a feature extractor and each iVector can be seen as a low-dimensional representation of the whole utterance.

This technique (*MultinIV*) can be used to model soft counts of high-dimensional, heterogeneous SNERFs as presented in Section 2.2. See [11] for further details of how all SNERFs can be represented using a single low-dimensional iVector and how the model parameters are trained using an iterative optimization scheme.

3.3. PLDA modeling of iVectors

The fixed-length iVectors extracted per utterance (from the *GaussIV* as well as from the *MultinIV* model) can now be used as input to a pattern recognition algorithm. Note that unlike in the standard JFA, where two subspaces are used to account for speaker and intersession variability, the iVector variant uses a single subspace accounting for all the variability. Therefore, the extracted vectors $\mathbf{w}$ are not free of channel effect, and intersession compensation must be eventually considered during classification.

For speaker verification a PLDA model [12] has been proposed to provide a probabilistic framework for modeling speaker and intersession variability in the iVector space. Model parameters can be trained using an EM algorithm [13]. 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". This can be evaluated analytically, and scoring can be performed very efficiently as described in [14].

4. Experiments

This section describes the experimental setup and results for the individual prosodic systems and for the combination of these systems with each other and with a baseline cepstral system.

4.1. Data

Results are presented on the telephone core conditions of the NIST Speaker Recognition Evaluations 2008 [1] (*dev*) and 2010 [2] (*eval*). Trials involve English conversational speech recorded over various telephone channels. Our development set is based on the original NIST SRE 2008 evaluation set, but was extended to include about two orders of magnitude more impostor samples, to adjust for the new DCF point. It includes 1,154 target and 1,516,837 nontarget trials. Our evaluation set corresponds to the official extended condition 5 of NIST SRE 2010 and contains 7,169 target and 408,950 nontarget trials.

Training of background, subspace, and PLDA models is performed on data from Switchboard corpora as well as NIST SRE 2004 – 2006 corpora. This set includes 13,482 recordings from 752 male and 16,782 recordings from 963 female speakers.

4.2. Prosodic systems

Experiments are carried out to evaluate the performance of the iVector modeling approach for the simple DCT features. For both, the *GaussJFA* and the *GaussIV* systems, we extract 13-dimensional DCT contour features (1 duration, 6 pitch and 6 energy values) and train gender-dependent multivariate universal background models (UBMs) with 512 Gaussian components and diagonal covariances. The *GaussJFA* and the *GaussIV* models are trained using sufficient statistics extracted for all background data using the same UBMs. For the *GaussJFA* model we train 100-dimensional speaker subspace $\mathbf{V}$ and 50-dimensional channel subspace $\mathbf{U}$. For the *GaussIV* model we train 300-dimensional total variability subspace $\mathbf{T}$ on the same data. These subspace sizes were found optimal in earlier experiments. The *GaussJFA* model is evaluated directly by log-likelihood ratio using a fast scoring technique [18] followed by zt-norm. The extracted DCT iVectors for all background data are used to train a full rank PLDA model. The PLDA model is then used to evaluate the log-likelihood ratio for speaker trials. Figure 1 shows results for the two DCT-based systems (green markers). The *DCT-GaussIV* system with PLDA (square) clearly outperforms the *DCT-GaussJFA* system (triangle) on all operating points on both test sets.

To compare the simple *DCT-GaussIV* system with the best prosodic system presented so far [15], we train a *SNERF-MultinIV* system on the same setup. The SMM models an ensemble of 1,638 multinomial distributions representing 9 different n-gram tokens of 182 individual SNERFs. We obtain 300 dimensional iVectors. While the *SNERF-MultinIV* system (blue diamonds in Figure 1) is still superior on both test sets for EER and old DCF, we achieve better results with the *DCT-GaussIV* system on both test sets in terms of new DCF.

As both prosodic systems perform very well, but are significantly different in terms of features as well as modeling approach, a combination of both seems natural. Since both modeling techniques translate the long-temporal prosodic feature vectors of variable size to a single fixed-length feature vector per utterance (what we call iVector), it is possible to simply concatenate the iVectors resulting from these diverse models and to model them jointly with a PLDA model. We train a single full-rank PLDA model on 600-dimensional iVectors. The effectivity of the joint modeling of complementary iVectors can be observed in Figure 1. The combination of *DCT-GaussIV* and *SNERF-MultinIV* iVectors (cyan hexagons) results in significant improvement over the best individual system on all operating points on both test sets, achieving an EER of 5.4% and a new DCF of 0.72 on 2008 data, which are (to our knowledge) the best results reported for a purely prosodic system.

4.3. Combination with cepstral baseline system

Our baseline system is a cepstral iVector system followed by a PLDA model (*CEP-GaussIV*). This system was the best-performing individual system from the ABC NIST SRE 2010 submission [16]. It is based on 60-dimensional cepstral features and a 2048-component full covariance UBM. Four hundred-dimensional iVectors are used and the dimension is further reduced to 200 by standard LDA and normalized by their length[1] before PLDA modeling. The first row of Table 1 gives the results for our two data sets[2].

Again, the iVector nature of our baseline system allows us to use a novel way of combining low- and high-level systems by simple concatenation of their iVectors and joint PLDA modeling. First, we apply an LDA reduction to 200 dimensions and length normalization to both 300-dimensional sets of prosodic iVectors. In this way we have three same sized sets of 200 dimensional iVectors (one cepstral and two prosodic). Next, we concatenate the cepstral iVectors separately with each of our prosodic iVectors to obtain two sets of four hundred-dimensional iVectors. Then we train a standard PLDA model with full rank of 400 for each type of combination. The second and third row of Table 1 give the results for these combinations. We see that we can achieve significant improvements for both *iVector fusions* of cepstral and prosodic features. Finally, we concatenate all three iVector types (one cepstral and two prosodic) and train a PLDA model with full rank of 600. The fourth row of Table 1 gives the results for this combination. We achieve further improvements leading to reductions as high as 21% relative on the challenging new DCF measure.

As a last experiment we compare this approach to the conventional score-level fusion. For this purpose we train a linear logistic regression [19] to fuse the three individual system scores on the development set and apply this fusion to the evaluation set. The last row of Table 1 indicates that consistent gains are also achieved by score-level fusion (as high as 13% on new DCF), but joint PLDA training of concatenated iVectors remains superior. iVector fusion of the cepstral system and the simple prosodic *DCT-GaussIV* system already outperforms the score-level fusion of all three systems.

5. Conclusions and Lookout

We present the first results on the use of total variability modeling of the mean supervector space for a set of prosodic features. We show that this iVector approach outperforms the standard JFA approach originally proposed for these features. We note that this improvement over JFA is observed only when the iVectors are modeled using the PLDA back end. No gain was observed during SRE 2010 system development [16] when iVectors were modeled with simpler scoring techniques [6].

[1] This pre-processing of iVectors is very helpful for cepstral iVectors but did not show any improvement for our prosodic iVectors

[2] We are aware that better results are reported in the literature, simply by training the PLDA on more data, which we did not have for SNERFs.

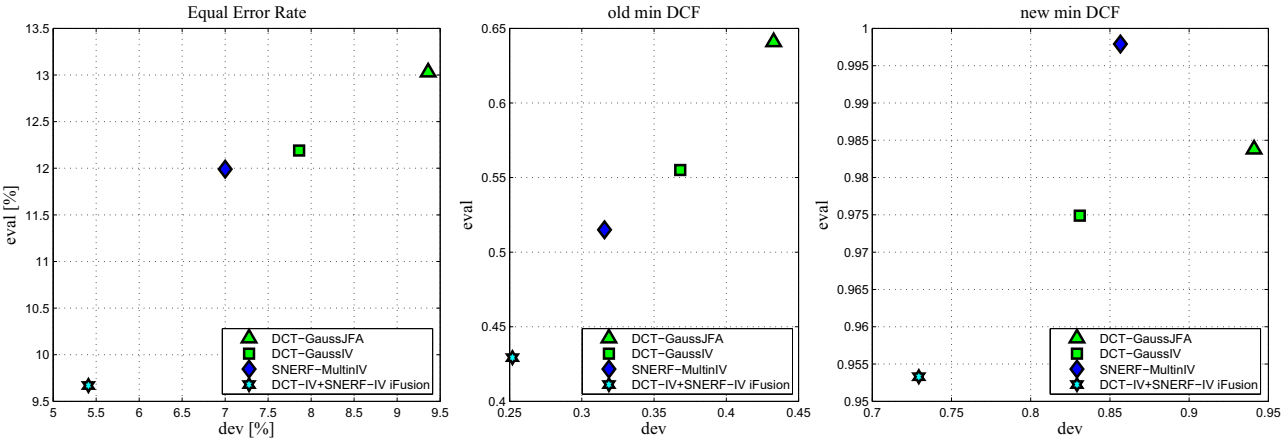

Figure 1: Results for SRE 2008 (dev) versus SRE 2010 (eval) in terms of EER, old DCF and new DCF, from left to right, for three different prosodic systems and combination of the two best.

System	DEV SRE 2008			EVAL SRE 2010		
	EER	old DCF	new DCF	EER	old DCF	new DCF
Cepstral iVector system *CEP-GaussIV*	2.02	0.090	0.471	3.14	0.155	0.504
Concatenated *CEP-GaussIV + DCT-GaussIV*	1.69	0.080	0.400	2.72	0.136	0.431
Concatenated *CEP-GaussIV + SNERF-MultinIV*	**1.65**	0.080	0.389	2.74	0.134	0.444
Concatenated *CEP-GaussIV + DCT-GaussIV + SNERF-MultinIV*	1.70	**0.075**	**0.368**	**2.63**	**0.129**	**0.421**
Score fusion *CEP-GaussIV + DCT-GaussIV + SNERF-MultinIV*	1.92	0.078	0.406	3.09	0.149	0.447

Table 1: Results for single cepstral baseline system (CEP-GaussIV) and for combinations with one or two prosodic iVector systems.

Furthermore, we present combination results of two prosodic systems, one where iVectors based on GMMs are used to model simple DCT features extracted from uniform regions and another one where iVectors based on multinomial distributions are used to model a complex set of syllable-level features. These two systems are different at both the feature and modeling levels. We show gains on the order of 20% when combining these two systems with respect to the single best. The combination is performed using an iVector-level fusion: the individual iVectors for the two systems are concatenated and the joint iVector is modeled using PLDA. An important advantage of iVector-level fusion compared to score-level fusion is that it can make use of the full information encoded in the iVectors while for the score-level fusion all information is already reduced to a single number.

The iVector-level fusion technique followed by PLDA modeling can also be applied to fuse heterogeneous features, such as low-level cepstral and high-level prosodic features. Using this procedure we achieve 20% relative improvement on new DCF over a cepstral iVector baseline, significantly outperforming score-level fusion. These are, to our knowledge, the largest relative gains obtained in speaker recognition from combination of cepstral systems with prosodic features in several years.

6. References

[1] NIST, "The NIST year 2008 speaker recognition evaluation plan," 2008. [Online]: http://www.itl.nist.gov/iad/mig//tests/sre/2008

[2] ——, "The NIST year 2010 speaker recognition evaluation plan," 2010. [Online]: http://www.itl.nist.gov/iad/mig//tests/sre/2010

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

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

[5] C.-Y. Lin *et al.*, "Language identification using pitch contour information," *Proc. ICASSP 2005, Philadelphia, PA*, pp. 601–604, 2005.

[6] N. Dehak *et al.*, "Modeling prosodic features with joint factor analysis for speaker verification," *Audio*, Jan 2007.

[7] M. Kockmann *et al.*, "Investigations into prosodic syllable contour features for speaker recognition," *Proc. of ICASSP, Dallas*, pp. 1–4, Sep 2010.

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

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

[10] N. Dehak *et al.*, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language processing*, pp. 1–23, Jul 2009.

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

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

[13] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Keynote presentation, Odyssey*, 2010.

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

[15] M. Kockmann *et al.*, "Recent progress in prosodic speaker verification," in *Proc. ICASSP, Prague*, 2011.

[16] N. Brummer *et al.*, "ABC system description for NIST SRE 2010," in *Proc. NIST 2010 Speaker Recognition Evaluation*. Brno University of Technology, 2010, pp. 1–20.

[17] L. Ferrer, "Statistical modeling of heterogeneous features for speech processing tasks," Ph.D. dissertation, Stanford University, 2009.

[18] O. Glembek *et al.*, "Comparison of scoring methods used in speaker recognition with joint factor analysis," *Proc. of ICASSP, Taipei*, 2009.

[19] E. de Villiers *et al.*, "BOSARIS toolkit," 2010. [Online]: http://sites.google.com/site/bosaristoolkit