A LINGUISTICALLY-MOTIVATED SPEAKER RECOGNITION FRONT-END THROUGH SESSION VARIABILITY COMPENSATED CEPSTRAL TRAJECTORIES IN PHONE UNITS

Joaquin Gonzalez-Rodriguez [1,2], J. Gonzalez-Dominguez[2], J. Franco-Pedroso[2] and D. Ramos[2]

[1] International Computer Science Institute, Berkeley, CA, USA
[2] ATVS, Universidad Autonoma de Madrid, Spain

ABSTRACT

In this paper a new linguistically-motivated front-end is presented showing major performance improvements from the use of session variability compensated cepstral trajectories in phone units. Extending our recent work on temporal contours in linguistic units (TCLU), we have combined the potential of those unit-dependent trajectories with the ability of feature domain factor analysis techniques to compensate session variability effects, which has resulted in consistent and discriminant phone-dependent trajectories across different recording sessions. Evaluating with NIST SRE04 English-only 1s1s task, we report EERs as low as 5.40% from the trajectories in a single phone, with 29 different phones producing each of them EERs smaller than 10%, and additionally showing an excellent calibration performance per unit. The combination of different units shows significant complementarity reporting EERs as 1.63% (100xDCF=0.732) from a simple sum fusion of 23 best phones, or 0.68% (100xDCF=0.304) when fusing them through logistic regression.

Index Terms— Speaker recognition, linguistic units, temporal trajectories, session variability, feature compensation.

1. INTRODUCTION[1]

Speaker recognition has been largely dominated in the last two decades by acoustic/spectral approaches both in terms of recognition accuracy and computational efficiency [9]. From the early GMM-UBM and SVM to the recent i-vector [3] and PLDA [8], once short-term linguistic-independent feature-vectors are extracted the relation between them and their associated linguistic information is lost. However, there is a large corpus of research in high-level speaker recognition [13], where the different linguistic information embedded into the speech signal are exploited in order to obtain better speaker characterization, especially when properly combined with acoustic systems. The added advantage of those low-plus-high-level systems over acoustic/spectral systems, critical a decade ago [11] even at the expense of much higher computational complexity, has largely vanished with the advent of highly efficient approaches as i-vectors and PLDA. Fortunately, recent contributions have successfully exploited the best of both approaches combining phonetic and prosodic conditioning to frame selection and UBM development [5] or unit-dependent prosodics [10] into state-of-the-art systems.

This work is an attempt to provide new linguistically-motivated feature vectors that can be directly exploited into state-of-the-art systems. Originated as a natural extension of our recent work on temporal contours in linguistic units (TCLU) [4] with formants, where we exploited the formant and formant bandwidth dynamics present in different types of linguistic units (phones, diphones, triphones, center phone in triphones, syllables and words), MFCC trajectories seemed a promising candidate towards better performance. However, the observed performance was not better than that with formant trajectories, as MFCCs are well-known to be seriously degraded by session variability. In an attempt to combine the known potential of temporal contours in linguistic units with the ability of factor analysis to deal with session variability compensation, a new front-end was developed providing linguistically-motivated feature vectors from non-uniform-length segments of session variability compensated cepstral temporal trajectories in phone units.

The remainder of the paper is organized as follows. In Section 2 we review factor analysis and its application into the feature domain, while in Section 3 we describe the proposed linguistically motivated front-end. Sections 4 and 5 describe the experimental protocol and system in use. Section 6 shows results for a variety of conditions and combinations of the available units, to finally conclude in Section 7 summarizing the main contributions and future extensions of this work.

2. FEATURE DOMAIN SESSION VARIABILITY COMPENSATION

State-of-the-art factor analysis techniques in speaker recognition have shown to properly address the problem of session variability but using spectral only systems we lose any reference to the linguistics or the temporal dynamics present in the speech signal. In order to take advantage of the potential of factor analysis for session variability compensation but keeping the discriminant information present in the temporal contours in linguistic units, we will use a factor analysis based compensation scheme in the feature domain as proposed in [14].

We follow here the FA assumption where a GMM means supervector of a speaker s and utterance h, μ_{sh}, is composed as the sum of speaker and session components as

$$\mu_{sh} = \mu_s + Ux_h$$

where the second term, formed by the session variability subspace U and *channel factors* associated to utterance h, x_h, is considered to be independent of the speaker s; then a feature domain session-variability compensation is performed based on subtracting the

[1] Supported by MEC grant PR-2010-123, MICINN project TEC09-14179, ForBayes project CCG10-UAM/TIC-5792 and Catedra UAM-Telefonica. Thanks to ICSI for hosting the preliminary part of this work.

corresponding additive session component to each observation vector, $o_h(t)$, as follows

$$\hat{o}_{h(t)} = o_h(t) - \sum_c \gamma_c(t) U_c x_h$$

being $\hat{o}_h(t)$ the resulting session variability compensated observations, $\gamma_c(t)$ the Gaussian occupation probability of frame t respect to the c-th UBM Gaussian component and U_c the sub-matrix of the session variability subspace corresponding to Gaussian c. To alleviate this costly operation, the sum in c uses to be constrained to the five most likely Gaussians per frame.

This and other similar approaches acting in the feature domain (i.e. feature latent factor analysis [2]), even though they involve a costly frame-by-frame compensation, have the prime advantage of allowing the use of any type of subsequent modeling schemes or classifiers once 'clean' features are obtained. This point makes it especially suitable for the front-end proposed in this work.

3. UNIFORM FEATURE EXTRACTION FROM VARIABLE LENGTH PHONE SEGMENTS

The temporal dynamics of speech have been used from the simplest (and successful) use of the spectral coefficients velocity (delta) and acceleration (delta-delta) to modulation spectrograms, frequency modulation features or even TDCT features (temporal DCT) (see [9] for a review). However, to the best of our knowledge none of the previous approaches, with the exception of SNERFs [12] and [10] for prosodic information, take advantage of the linguistic knowledge provided by an automatic speech recognizer to extract non-uniform-length sequences of spectral vectors to be converted into constant-size feature vectors characterizing the temporal-spectral information in a given phone.

In our proposed front end, once the original sequence of MFCC vectors have been session variability compensated, we obtain a constant-size feature vector from non-uniform-length phonetic segments as in our previous work in [4] with formant trajectories. In this case, from the phone labels provided by SRI-Decipher [6], the trajectories in segments of varying length of 19 static and 19 delta session variability compensated MFCC are extracted. As the shape of the trajectories in a given phone should be equivalent independently of the phone duration, all repetitions are duration equalized for subsequent trajectory coding with a fifth order discrete cosine transform. In this sense, the entire session variability compensated MFCC feature vectors belonging to a given phone repetition (a variable size matrix of size 38 x #frames/phone) are compressed into a fixed size vector of P coefficients per trajectory. In our experiments, the resulting feature vector per phone has either size 95 (19MFCC -or deltaMFCC- x 5 DCT coefs/trajectory) or 190 (19MFCC+19deltaMFCC=38 x 5).

4. SYSTEM DESCRIPTION

In order to properly evaluate the individual performance of every unit under analysis with the proposed front-end, we have developed a system able to produce calibrated likelihood ratios (LR) for every individual unit in a manner similar to that fully detailed in [4]. As the number of repetitions of the 41 phones under analysis in an utterance varies almost linearly among them from 10 to 160, we need to be able to compute LRs from very limited amounts of data. A MultiVariate Likelihood Ratio (MVLR) technique as MVK (MV Kernel) [1], well known in some forensic disciplines, has been selected for producing calibrated LRs directly from the observed data, without the need for additional data to train score to LR converters. We use it here in an identical way to our previous work with formants, so details can be found in [4]. This type of systems directly producing calibrated LRs per linguistic unit of analysis have a double interest, as individual unit LRs can directly be interpreted by humans (e.g. for forensic reporting or linguistic analysis) or machines, but also be further combined into a single LR per trial either from rule-based fusion or, when additional calibration data in equivalent conditions is available, data-trained fusion as logistic regression.

5. DATASETS AND EXPERIMENTAL SETUP

The experimental protocol is identical as the one detailed in our recent work on formants [4], which can be summarized in the use of the NIST SRE 2004 English-only 1s1s trials and data, which comprises both native and nonnative speakers across 9,655 same-sex different-telephone-number trials from 208 speakers (123 female and 85 male). All reported TCLU systems elicit likelihood ratios, so C_{llr} and *minCllr* (and its difference, calibration loss) are used to evaluate the goodness of the different detectors. For every trial, the data and unit's LRs from speakers different from those in the trial are used for MVK background modeling and logistic regression fusion. The gender dependent U matrices for session variability compensation have been trained with PCA plus 5 EM iterations with telephone data from Switchboard I&II, NIST SRE05 and SRE06 (no SRE04 in U). Gender dependent UBMs, only used for feature compensation, are trained with 4 million feature vectors each from Switchboard I&II, SRE04, 05 and 06.

6. RESULTS

6.1. Cepstral versus formant contours

Our first objective was to improve the good results previously obtained fusing formant contours, based on individual EERs per unit in the 25-35% range. However, the results with highly promising MFCC contours (with CMN-RASTA and 3 seconds feature warping) were disappointing, probably because of the degradation introduced by session variability. Those results can be seen in the first two columns of table 1, where MFCC contours performed similarly or even worse than formant ones.

	EER (%)				
ph	FB123	MFCC	CFM	ΔCFM	CFM+ΔCFM
IY	31.72	38.67	6.73	7.46	5.40
IH	29.16	34.35	6.50	7.19	5.57
AH	28.53	34.62	6.21	8.19	5.62
AY	26.66	38.30	7.18	7.94	5.72
EH	27.85	34.11	7.18	7.07	5.73
K	33.95	35.90	7.04	7.88	5.84
S	37.88	40.87	7.18	6.39	6.01
P	30.57	20.01	7.18	8.76	6.02
OW	27.21	36.77	6.97	8.12	6.73
G	30.26	27.00	8.44	7.52	6.79

Table 1. Comparison of individual EERs in 10 best phone systems from trajectories of formants and bandwidths (FB123), MFCC, and session variability compensated MFCC (CFM and ΔCFM) in the SRE04 English 1side1side task.

However, when feature domain session variability compensation is performed over the MFCC coefficients, the new trajectories show a significant performance gain when computed in static (19 channel compensated MFCC, denoted CFM), delta (19 ΔCFM) and static plus delta configurations (19CFM+19ΔCFM), suggesting highly consistent and discriminant trajectories in phonetic units after feature domain factor analysis session variability compensation.

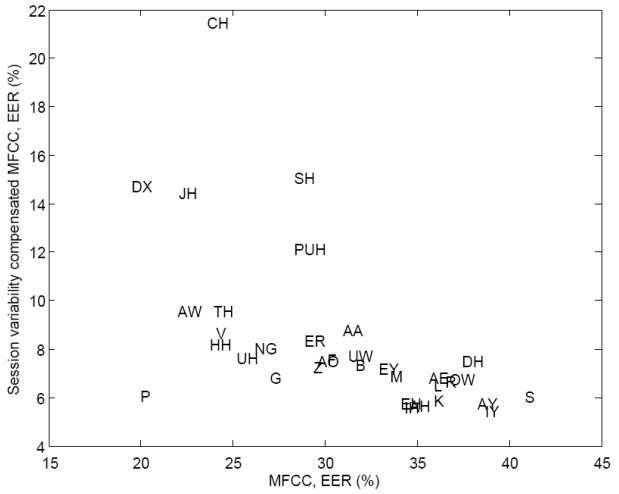

Figure 1. Comparison of EERs from trajectories of MFCCs (with CMN-Rasta-Warping) versus trajectories of session variability compensated MFCCs with 34 different phones.

phone	EER (%)	DCF	Cllr	minCllr
AA	8.76	0.028	0.316	0.271
AE	6.80	0.021	0.213	0.202
AH	5.62	0.022	0.201	0.190
AO	7.49	0.028	0.277	0.249
AY	5.72	0.028	0.221	0.205
B	7.29	0.024	0.228	0.219
DH	7.49	0.025	0.260	0.239
EH	5.73	0.024	0.192	0.181
ER	8.30	0.031	0.265	0.252
EY	7.15	0.026	0.246	0.225
G	6.79	0.025	0.230	0.212
HH	8.18	0.027	0.283	0.261
IH	5.57	0.020	0.190	0.181
IY	5.40	0.023	0.193	0.181
K	5.84	0.024	0.209	0.198
M	6.84	0.022	0.244	0.219
NG	8.02	0.032	0.287	0.261
OW	6.73	0.025	0.240	0.216
P	6.02	0.021	0.207	0.195
S	6.01	0.021	0.202	0.192
UW	7.70	0.031	0.290	0.249
V	8.65	0.036	0.301	0.276
Z	7.19	0.025	0.236	0.222

Table 2. Individual discrimination and calibration performance of each of 23 best phones in the SRE04 English 1side1side task with channel compensated MFCC trajectories.

6.2. Session compensated cepstral contours

Session variability compensated cepstral contours per phone show an excellent discrimination performance (as observed in table 2 for 23 best phones), where 29 out of the 41 phones under analysis obtain EERs smaller than 10% (fig. 1). But especially remarkable is the excellent calibration shown by all the LRs elicited by all those systems, as shown by the very limited calibration loss observed (difference between columns 4 and 5 in table 2).

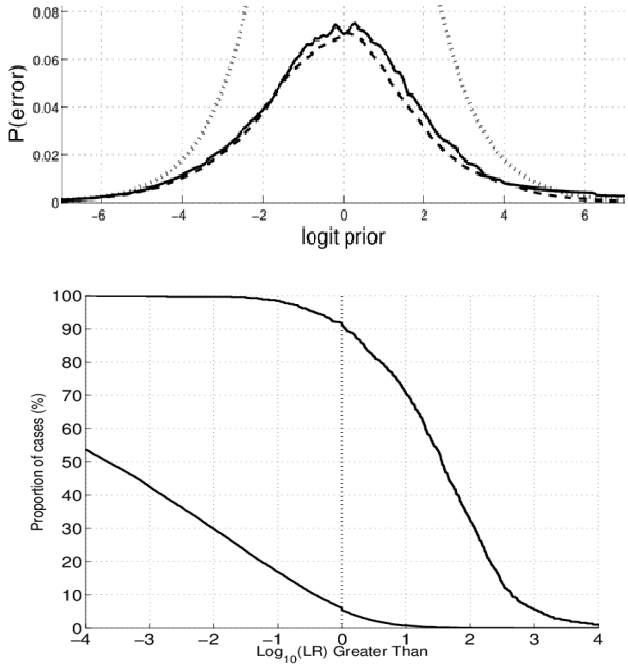

Figure 2. APE and Tippet plot of elicited LRs from a single phone system (phone IY) in the SRE04 English 1side1side task with session variability compensated MFCC trajectories.

In higher detail, we show in figure 2 a graphical illustration of the goodness of the calibration of the elicited LRs for a sample phone system for every application prior in the APE plot, and low rates of misleading evidence (targets with LR<1 and non-targets with LR>1) in the Tippet plot, especially if we recall that those results are produced for the SRE04 1s1s task from the analysis of a single phone. Having a good calibration, without the need for additional calibration data (LRs were directly obtained with MVK from the observed and background data), the information provided by linguistic unit LRs are directly interpretable by humans or machines, keeping intact their potential for fusion, as shown in the following section.

6.3. Fusion of session compensated cepstral contours

Different types of fusion have been performed in order to check the complementarity of the discrimination abilities of the different units under analysis. First, a rule-based fusion has been performed with N-best phones (N=23) through simply averaging the unit dependent LRs from Section 6.2. Secondly, as lots of trials from different speakers are available in similar conditions (those of SRE04 data), a data-trained fusion through logistic regression has also been performed with the protocol described in Section 5. The

results are excellent with both fusion approaches, with EERs in the SRE04 task as low as 1.63% with sum fusion, or 0.68% with logistic regression, obtaining in the latter case an even better combined result at the expense of the need for additional data to train the score to likelihood ratio converter.

fusion	gender	EER (%)	DCF
Sum	Male	2.98	0.01028
	Female	0.70	0.00289
	M+F	1.63	0.00732
logreg	Male	1.10	0.00266
	Female	0.30	0.00140
	M+F	**0.68**	0.00304

Table 3. Performance of different fusions of 23 best phones with session variability compensated MFCC (CFM+ΔCFM) trajectories in the SRE04 English 1side1side task.

Finally, in table 4 we compare the performance of different spectral and TCLU systems. Especially remarkable is the comparison of the FAu50 system, which is a feature domain compensated factor analysis raw (no score normalization) spectral system, with the channel factor compensated TCLU systems. All of them use exactly the same compensated features as input but the TCLU systems additionally exploit the temporal structure of those compensated features within each of the linguistic units in use, providing a significant performance improvement.

	#units	EER (%)	DCF
GMM-MAP	-	14.01	0.05958
FAu50	-	4.25	0.01995
Formants	79	3.88	0.01940
Sum CF_MFCC	23	1.63	0.00732
LogReg CF_MFCC	23	0.68	0.00304

Table 4. Performance comparison of two raw spectral-only systems and three TCLU systems in the SRE04 English 1s1se task.

We are aware that the reported error rates of TCLU systems, obtained in a task with 208 speakers but only 9655 trials, have been obtained in the fused systems from a very small number of errors, which forces us to extend as soon as possible this work to more recent and challenging SRE tasks. In any case, special care has been taken in the experimental protocol to ensure that for every trial the system had no knowledge (background data, training scores, channel conditions) about the speakers involved in the trial, and the reported results are then realistic for the given task.

7. SUMMARY AND CONCLUSIONS

In this paper a new linguistically motivated front end for speaker recognition has been introduced. The combined use of feature domain factor analysis to deal with session variability without losing the temporal dynamics in the speech signal, and TCLU systems to exploit the temporal structure within linguistic units, provide a significant improvement in performance. Comparing with the improvements obtained including unit dependent prosodics into i-vector and PLDA systems [10], the proposed front end turns into a promising candidate to be included in this type of systems because of the reported exceptional performance of compensated cepstral contours.

8. REFERENCES

[1] Aitken, C. G. G. and Lucy, D., "Evaluation of trace evidence in the form of multivariate data", Applied Statistics 53, pp. 109-122, with corrigendum pp. 665-666, 2005.

[2] Campbell, W.M., et al, "A Comparison of Subspace Feature-Domain Methods for Language Recognition", Proceedings of Interspeech 2008, September 2008.

[3] Dehak, N., et al., "Front-End Factor Analysis for Speaker Verification", IEEE Trans. on Audio, Speech and Lang. Proc., 19(4), 788-798, May 2011.

[4] Gonzalez-Rodriguez, J., "Speaker recognition using temporal contours in linguistic units: the case of formant and formant-bandwidth trajectories", Interspeech 2011, pp. 133-136, Florence, Italy, 2011.

[5] Hewlett, M., Ferrer, L., Shriberg, E. and Stolcke, A., "Constrained Cepstral Speaker Recognition Using Matched UBM and JFA Training", Proc. Interspeech 2011, pp. 141-144, Florence, Italy, 2011.

[6] Kajarekar, S. et al., "The SRI NIST 2008 Speaker Recognition Evaluation System", Proc. IEEE ICASSP'09, pp. 4205-4209, Taipei, 2009.

[7] Kenny, P. et al., "A Study of Interspeaker Variability in Speaker Verification", IEEE Trans. on Audio, Speech and Lang. Proc., 16(5):980–988, 2008.

[8] Kenny, P., "Bayesian speaker verification with heavy tailed priors", Keynote presentation at Odyssey 2010, Brno, 2010.

[9] Kinnunen, T., and Li, H., "An overview of text-independent speaker recognition: from features to supervectors", Speech Communication, vol. 52, pp. 12-40, 2010.

[10] Kockmann, M., Ferrer, L., Burget, L., Cernocky, J., "iVector Fusion of Prosodic and Cepstral Features for Speaker Verification", Proc. Interspeech 2011, pp. 265-268, Florence, Italy, 2011.

[11] Reynolds, D. et al., "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition", Proc. Icassp 2003, vol. IV, pp. 784-7, 2003.

[12] Shriberg, E., "Modeling prosodic feature sequences for speaker recognition", Speech Communication, 46 (3-4), July 2005, pp. 455-472, Jan. 2005.

[13] Shriberg, E., "Higher-level features in speaker recognition", in Speaker Classification I: Fundamentals, Features and Methods, C. Müller, Ed., Springer LNCS 4343, pp. 241-259, Springer, 2007.

[14] Vair, C. et al., "Channel Factors Compensation in Model and Feature Domain for Speaker Recognition", Proc. Odyssey Speaker and Language Recognition Workshop 2006, pp. 1- 6, 2006.