Using Phone Log-Likelihood Ratios as Features for Speaker Recognition

Mireia Diez, Amparo Varona, Mikel Penagarikano, Luis Javier Rodriguez-Fuentes, Germán Bordel

GTTS, Department of Electricity and Electronics, ZTF/FCT
University of the Basque Country UPV/EHU, Barrio Sarriena, 48940 Leioa, Spain
mireia.diez@ehu.es

Abstract

The so called Phone Log-Likelihood Ratio (PLLR) features, computed on phone posterior probabilities provided by phonetic decoders, convey acoustic-phonetic information in a sequence of frame-level vectors. Thus, PLLRs can be easily plugged into traditional acoustic systems just by replacing MFCCs, PLPs or whatever other representation. PLLR features were used under an iVector-PLDA approach in our submission to the NIST 2012 Speaker Recognition Evaluation (SRE). In this work, we present a report of the goodness of these features for speaker recognition. Results on the telephone clean speech condition of the NIST 2010 and 2012 SRE show that, although the system based on PLLR features does not reach state-of-the-art performance, including it in a fusion with a traditional acoustic based system (trained on MFCC features) provides remarkable gains in performance (among the best reported in the NIST 2012 SRE telephone without added noise condition), revealing a fruitful way of using acoustic-phonetic information for speaker recognition.

Index Terms: Speaker Recognition, Phone Log-Likelihood Ratios, NIST 2012 SRE, iVectors, Probabilistic Linear Discriminant Analysis

1. Introduction

Speaker Recognition (SR) refers to the task of recognizing a person given an audio signal containing his/her voice. Extensive overviews of actual feature extraction techniques and modeling approaches for SR can be found in [1] and [2]. Most SR systems are based on short-term spectral low-level features, which model vocal-tract properties and the spectral envelope of the sounds [1]. Among them, the widely used Mel Filter Cepstral Coefficients (MFCC) stand out as one of the most popular. Others, like Perceptual Linear Prediction Coefficients (PLP) or Linear Predictive Cepstral Coefficients (LPCC), are also successfully applied in several approaches [3], [4], [5].

Though short-term spectral features seem to be the most discriminative for SR, many works have explored the use of different higher-level features which could capture speaker-specific linguistic and behavioral aspects not reflected at the spectral level [1], [2]. For example, *voice source* features, which take into account the glottal excitation signal of voiced sounds [6] or *prosodic features* which characterize non-segmental aspects of speech (syllable stress, intonation patterns, speaking rate and rhythm) [7]. Other popular high-level approaches are based on the use of phone/word n-grams obtained from phonetic/word decoders [8]. A variant is proposed in [9], making use of phone n-grams obtained from the segments corresponding to the most frequent words. In general, all these studies found that, though non-spectral features might not be as discriminative as spectral ones, they do provide complementary information that helps improving system performance when different approaches are fused [10].

Regarding speaker modeling in the SR field, nowadays, the *Total Variability Factor Analysis* approach, which extracts low-dimensional features known as *iVectors*, has become state-of-the-art, due to its excellent performance, low complexity and low dimensionality [11] [12]. iVectors are then processed under different modeling approaches, such as variants of Probabilistic Discriminant Analysis (PLDA), which stand out as the current trend for SR systems [13], [14] outperforming classical cosine-distance scoring of iVectors with normalization [11]. PLDA was among the most used techniques in the last NIST 2012 SRE evaluation [3], [15].

The iVector PLDA approach was also applied in our submission (EHU) for the last NIST 2012 SRE evaluation.When searching for an alternative set of features to the standard MFCCs, we came to the idea of using Phone Log-Likelihood Ratios (PLLRs). PLLRs are computed from phone posterior probabilities provided by phone decoders, conveying high level phonetic and acoustic short-term information into frame-level vectors, which allows plugging the features directly into any short-term spectral feature based system (e.g. iVector-PLDA). PLLRs had been previously used successfully for Spoken Language Recognition with remarkable results [16].

In this paper, we present a report of the utility of PLLR features for SR: a detailed description of the PLLR features is given and sets of experiments carried out on both NIST 2010 and 2012 SREs, on the telephone clean speech conditions, are reported. PLLR-based systems are compared to (and fused with) state-of-the-art MFCC-based baseline systems, revealing a complementarity that enhances significantly system performance.

The rest of the paper is organized as follows: Section 2 defines the PLLR features. Section 3 describes the iVector PLDA approach used in the SR systems. The experimental setup is described in section 4 and results are shown and discussed in Section 5. Finally, conclusions are outlined in Section 6.

2. PLLR Features

2.1. Definition of PLLR features

Given a phone decoder with a set of N phone units, each of them represented typically by means of a model of S states, we assume that the acoustic posterior probability of each state s $(1 \leq s \leq S)$ of each phone model i $(1 \leq i \leq N)$ at each frame t, $p(i|s,t)$, is directly provided by the phone decoder. Then, the acoustic posterior probability of a phone unit i at each frame t

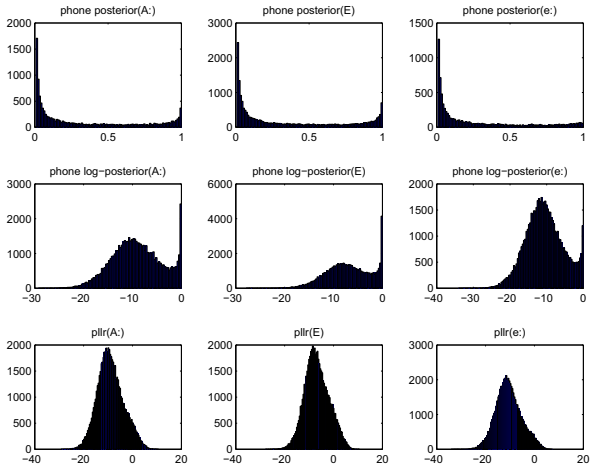

Figure 1: Distributions of frame-level posteriors (first row), phone log-posteriors (second row) and phone log-likelihood ratios (third row) for three Hungarian phones (A:, E and e:).

can be computed by adding the posteriors of its states:

$$p(i|t) = \sum_{\forall s} p(i|s,t) \tag{1}$$

The phone posteriors computed according to Eq. 1 exhibit extremely skewed and sparse (non-Gaussian) distributions (see Figure 1, first row). This is not suitable for the kind of models we apply in SLR systems, since they typically assume that features are Gaussian-distributed. To address the non-Gaussian nature of phone posteriors, we can transform them into log-posteriors, obtaining more suitable but still non-Gaussian distributions (see Figure 1, second row). If we go further and take the logarithm of the likelihood ratio, the obtained distributions are nearly Gaussian (see Figure 1, third row):

$$LLR(i|t) = \log \frac{p(i|t)}{\frac{1}{(N-1)}(1-p(i|t))} \qquad i = 1, ..., N \tag{2}$$

In this way, we obtain an N-dimensional feature vector per frame t, carrying the same information as the N phone posteriors, but featuring approximately Gaussian distributions. These are the Phone Log-Likelihood Ratio (PLLR) features used in this work.

2.2. Why PLLRs may be suitable for Speaker Recognition

Speech sounds are produced in a different way by each speaker. A phone decoder can be seen as a reference system for representing the speech sounds of any speaker in terms of the activation of its phonetic units. This seemingly simple representation involves a complex model, since each phonetic unit gets activated for specific sequences of spectral features, which include both static and dynamic information. Some units could be strongly correlated among each other, so they would get activated at the same time (e.g. all the nasal consonants would activate when a nasal sound appeared in the input), but each phonetic unit would also provide specific information that would help catching the subtle differences between sounds uttered by each speaker. In fact, the goodness of this representation depends on the richness of the inventory of phonetic units (i.e. how well they cover the sounds of each speaker) and the optimality of the classification/mixing process (i.e. how well the phone decoder estimates the activation of phonetic units that best represents any sound). On the other hand, since the phone decoder is trained on a large and diverse dataset, the PLLR features are expected to be robust to channel and other sources of variability.

In summary, PLLRs may be seen just as an alternative way of representing speech sounds, including the subtle differences in the realization of sounds among speakers, which makes them suitable for SR.

3. The iVector PLDA Approach

Under the i-Vector modeling assumption, an utterance GMM supervector (stacking GMM mean vectors) is defined as [11][12]:

$$M = m + Tw \tag{3}$$

where M is the utterance dependent GMM mean supervector, m is the utterance independent mean supervector, T is the total variability matrix (a low-rank rectangular matrix) and w is the so called i-Vector (a normally distributed low-dimensional latent vector). That is, M is assumed to be normally distributed with mean m and covariance TT^t. The latent vector w can be estimated from its posterior distribution conditioned to the Baum-Welch statistics extracted from the utterance with regard to a background GMM (Universal Background Model, UBM). The i-Vector approach maps high-dimensional input data (a GMM supervector) to a low-dimensional feature vector (an i-Vector), hypothetically retaining most of the relevant information.

Under the Gaussian PLDA approach, an observation j (in this case, an iVector) of speaker i, is supposed to be modeled by:

$$w_{ij} = Vy_i + Ux_{ij} + z_{ij} \tag{4}$$

with:

$$y_i \sim \mathcal{N}(0, \mathbf{I}) \tag{5}$$
$$x_{ij} \sim \mathcal{N}(0, \mathbf{I}) \tag{6}$$
$$z_{ij} \sim \mathcal{N}\left(0, \mathbf{D^{-1}}\right) \tag{7}$$

where D is a diagonal precision matrix and the hidden variables y_i and x_{ij} are the speaker and channel factors, respectively [17], while z_{ij} is the noise term accounting for the rest of the variability. The model $\mathcal{M} = (V, U, D)$ is estimated by EM.

4. Experimental setup

4.1. MFCC Feature Extraction

MFCC features were computed in frames of 25 ms at intervals of 10 ms, by means of the Sautrela toolkit [18]. The MFCC set comprised 13 coefficients, including the zero (energy) coefficient. Cepstral Mean Subtraction (CMS) [19] and Feature Warping [20] were applied to cepstral coefficients. An energy based VAD was applied, which removed frames with energies more than 30db below the maximum. Finally, the feature vector was augmented with dynamic coefficients (13 first-order and 13 second-order deltas), resulting in a 39-dimensional feature vector.

4.2. PLLR Feature Extraction

In this work, the open software Temporal Patterns Neural Network (TRAPs/NN) phone decoder for Hungarian, developed by the Brno University of Technology (BUT) [21], has been applied to compute PLLRs. The BUT decoder for Hungarian includes 61 phonetic units, each featuring a three-state model, which means that three posterior probabilities per unit are computed at each frame, encoded in the following way:

$$x(i|s,t) = \sqrt{-2\log p(i|s,t)} \tag{8}$$

Thus, the posterior probability $p(i|s,t)$ can be obtained as follows:

$$p(i|s,t) = e^{-\frac{(x(i|s,t))^2}{2}} \tag{9}$$

Before computing log-likelihood ratios, the three non-phonetic units of the BUT decoder for Hungarian: *int* (intermittent noise), *pau* (short pause) and *spk* (non-speech speaker noise), are integrated into a single non-phonetic unit model. Then, a single posterior probability is computed for each phone model i ($1 \leq i \leq N$), by adding the posterior probabilities of all the states in the corresponding model (Equation 1). Finally, log-likelihood ratios are computed according to Equation 2. In this way, 59 PLLR features are computed at each frame t. The feature vector is then augmented with first order deltas [16], resulting in a 118-dimensional feature vector.

Voice activity detection is performed by removing the feature vectors whose highest PLLR value correspond to the integrated non-phonetic unit.

4.3. iVector PLDA Configuration

4.3.1. Preprocessing

The Qualcomm-ICSI-OGI (QIO)[22] noise reduction technique (based on Wiener filtering) was independently applied to the audio streams. The full audio stream was taken as input to estimate noise characteristics, thus avoiding the use of voice activity detectors on which most systems rely to constrain noise estimation to non-voice fragments.

4.3.2. UBM

For each system, two gender dependent UBMs (the same for NIST 2010 and 2012 experiments), each consisting of 1024 mixture components, were estimated on the dataset used for the EHU NIST 2010 SRE submission (a subset of SRE04, SRE05 and SRE06 consisting of 4882 signals) [23], using the Sautrela toolkit.

4.3.3. iVector Extractor

Gender dependent Total Variability matrices were estimated for each system on each training set (for NIST 2010 and 2012 SRE), by means of Sautrela. According to experiments on NIST 2010 SRE data, the i-Vector dimensionality was fixed to 500.

For the NIST 2010 SRE, the Total Variability Matrix was trained on a subset of signals from SRE04 and SRE06, that amounted to 8.117 signals. In the case of NIST 2012 SRE, the *single_file_per_ldc_id_map* list was used for training the iVector total variability matrix. Note that speech signals are not repeated in that list. Additionally, 590 channel-balanced randomly chosen signals from the *Follow-Up* set were also used for training, in order to increase the number of microphone-channel signals in the training set. This training set consisted of 21.176 signals.

4.3.4. Gaussian PLDA

Gender dependent Gaussian PLDA systems [17] were estimated on the training sets of NIST 2010 and 2012 SRE. PLDA systems were estimated using a speaker subspace of size 150, a channel subspace of size 400 and 20 Expectation-Maximization (EM) iterations.

The PLDA systems for NIST 2010 SRE were trained on a subset of signals obtained from SRE04, SRE06 and SRE08 plus Follow up signals, which amounted to 23.302 signals. For NIST 2012 SRE, the PLDA systems were trained on the same set of signals as the Total Variability matrix (see Section 4.3.3)

4.3.5. LLR Estimation

PLDA system scores $s(t,u)$ were used as log-likelihoods. For NIST 2012 SRE, the likelihood of a test utterance u given a speaker i was computed as the average likelihood of u over all the training signals t of that speaker, as follows:

$$p(u|i) = \frac{1}{|Train(i)|} \sum_{t \in Train(i)} e^{s(t,u)} \tag{10}$$

Finally, speaker log-likelihood ratios were computed from speaker log-likelihoods using flat priors (note that the calibration does the offset correction to the flat prior).

4.4. Fusion and Calibration

Calibration and fusion were estimated and applied by means of the Bosaris toolkit [24]. As for the PLDA estimation, the whole training set was used to estimate the calibration/fusion parameters. To avoid over-optimistic target scores during the calibration/fusion estimation, the *self* PLDA scores $s(u,u)$ where excluded from the LLR computation.

For NIST 2010 SRE experiments, the performance on both, the *old* (P_{fa}=0.01, C_{miss}=10, C_{fa}=1) and *new* (P_{fa}=0.001, C_{miss}=1, C_{fa}=1) operating points was computed. For the NIST 2012 SRE, the *effective prior* was set to 0.001.

4.5. Test sets and Evaluation Measures

On the NIST 2010 SRE, experiments were carried out in the *Conversational Telephone Speech in Training and Test* condition. On the NIST 2012 SRE, the *Train on Multiple Segments, Test on Telephone with No Added Noise* condition was selected for experimentation.

Results are shown in terms of Equal Error Rate (EER), Minimum Detection Cost Function (MinDCF) and Actual Detection Cost Function (ActDCF), as defined by NIST for each SRE [25], [26].

5. Results

5.1. Results on NIST 2010 SRE

Table 1 shows results for the iVector-PLDA systems trained on MFCC features and PLLR features, and the fusion of both, for the 2008 and 2010 calibration points. Results show that the acoustic system clearly outperforms the PLLR based approach, but the latter gets yet quite good performance compared to those usually attained by phonotactic systems on SR tasks.

The fusion of both approaches attains 19% and 16% relative improvements in terms of MinDCF and ActDCF with regard to the MFCC based approach in the 2008 operating point. Results

Table 1: Results of PLLR and MFCC feature based approaches, and fusion of them on the NIST 2010 SRE telephone-telephone speech condition, for the 2008 and 2010 operating points.

System	Op. point	EER	MinDCF	ActDCF
MFCC	2008	4.39	0.199	0.210
	2010	4.42	0.606	0.652
PLLR	2008	7.80	0.351	0.359
	2010	7.75	0.763	0.774
Fusion	2008	3.74	0.162	0.176
	2010	3.77	0.527	0.588

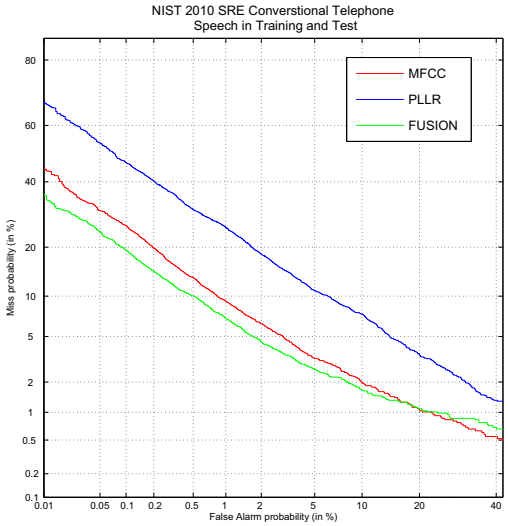

Figure 2: DET curves for the MFCC and PLLR based systems, and their fusion on the Conversational Telephone Speech in Training and Test condition of NIST 2010 SRE

show a miscalibration in the 2010 operating point, yet 13% and 10% relative improvements are attained in terms of MinDCF and ActDCF with regard to the acoustic approach.

Figure 2 shows the DET curves for the MFCC and PLLR based systems, and their fusion. Clear performance gains are observed in all operating points, with regard to the MFCC-based system, when both approaches are fused.

5.2. Results on NIST 2012 SRE

Results on the SRE 2012 dataset are shown in Table 2. Once again, the result attained by the acoustic system is better that the one obtained with the PLLR-based approach (0.296 vs 0.440 in terms of ActDCF).

However, the fusion of both approaches attains a 23% relative improvement in terms of MinDCF and a 19% relative improvement in terms of ActDCF with regard to the acoustic approach, which reveals a complementarity between the features.

Figure 3 shows the DET curves for the MFCC and PLLR based systems, and their fusion, with consistent performance gains in all operating points when both approaches are fused.

Table 2: Results of PLLR and MFCC feature based approaches, and fusion of them on the NIST 2012 SRE phone call with no added noise speech condition.

System	EER	MinDCF	ActDCF
MFCC	1.83	0.277	0.296
PLLR	3.12	0.419	0.440
Fusion	1.39	0.213	0.239

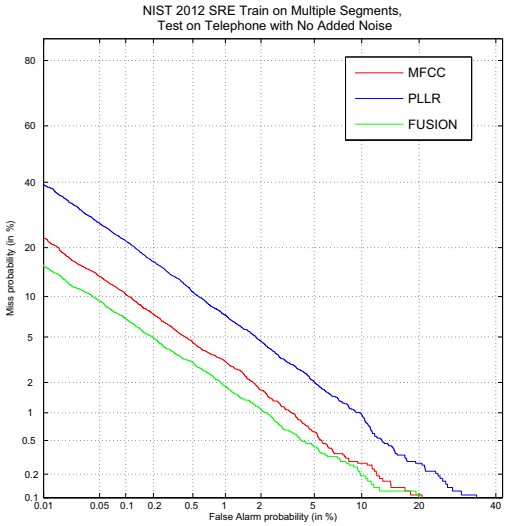

Figure 3: DET curves for the MFCC and PLLR based systems, and their fusion on the Train on Multiple Segments, Test on Telephone with No Added Noise condition of NIST 2012 SRE

6. Conclusions

In this work, Phone log-Likelihood Ratios have been used as features for Speaker Recognition tasks under an iVector PLDA approach. It has been shown that PLLRs can be easily plugged into short-term spectral feature based systems, by simply replacing the feature vector. Experiments have revealed that, though the PLLR-based system does not reach the performance of the baseline approach (relying on MFCC features under the same iVector-PLDA approach), it does provide significant gains in performance when both systems are fused: between 10% and 23% relative improvements, in terms in ActDCF, in the phone call clean speech core conditions of the NIST 2010 and 2012 SRE.

PLLR features, which convey acoustic and phonetic information in a short-term frame-level vector, provide a new and successful way of using high-level phonetic information for SR tasks.

7. Acknowledgments

This work has been supported by the University of the Basque Country, under grant GIU10/18 and project US11/06; and by the Government of the Basque Country, under program SAIOTEK (project S-PE12UN055); Mireia Diez is supported by a 4-year research fellowship from the Department of Education, University and Research of the Basque Country.

8. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010.

[2] Q. Jin and T. F. Zheng, "Overview of front-end features for robust speaker recognition," in *Proc. APSIPA*, 2011.

[3] N. Brummer, L. Burget, C. S., O. Glembek, J. A., M. Karafiat, K. P., P. Matejka, O. P., P. O., S. J., S. M., S. T., and A. Swart, "ABC System description for NIST SRE 2012," in *2012 NIST Speaker Recognition Evaluation (SRE) Workshop*, Orlando, USA, 11-12 December 2012.

[4] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210–229, 2006.

[5] T. Kinnunen, V. Hautamaki, and P. Franti, "Fusion of spectral feature sets for accurate speaker identification," in *SPECOM*, St. Petersburg, Russia, September, 2004, pp. 361–365.

[6] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *ICASSP*, 2008, pp. 4821–4824.

[7] M. Kockmann, L. Ferrer, L. Burget, and J. Cernocký, "ivector fusion of prosodic and cepstral features for speaker verification," in *INTERSPEECH*, 2011, pp. 265–268.

[8] W. Campbell, J. Campbell, T. Gleason, D. Reynolds, and W. Shen, "Speaker verification using support vector machines and high-level features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2085–2094, Sept.

[9] H. Lei and N. Mirghafori, "Word-conditioned phone n-grams for speaker recognition," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April, pp. IV–253–IV–256.

[10] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The supersid project: exploiting high-level information for high-accuracy speaker recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 4, April, pp. IV–784–7 vol.4.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[12] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *Proceedings of the Interspeech 2011*, Florence, Italy, August 27-31 2011, pp. 857–860.

[13] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV*, 2007, pp. 1–8.

[14] P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocký, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *ICASSP*, 2011, pp. 4828–4831.

[15] T. Hasan, S. O. Dasjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. L. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," in *Proceedings of ICASSP 2013*, Vancouver, Canada, May, 2012.

[16] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "On the Use of Log-Likelihood Ratios as Features in Spoken Language Recognition," in *IEEE Workshop on Spoken Language Technology (SLT 2012)*, Miami, Florida, USA, December 2012.

[17] N. Brummer, "The EM algorithm and Minimum Divergence applied to PLDA," Tech. Rep., 2010. [Online]. Available: https://sites.google.com/site/nikobrummer

[18] M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework," in *Proceedings of the ASRU Workshop*, San Juan, Puerto Rico, December 2005, pp. 386–391.

[19] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Acoustical Society of America Journal*, vol. 55, pp. 1304–1312, 1974.

[20] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, 2001, pp. 213–218.

[21] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.

[22] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *Proceedings of ICSLP2002*, 2002.

[23] M. Penagarikano, A. Varona, M. Diez, L. J. Rodriguez-Fuentes, and G. Bordel, "University of the Basque Country System for NIST 2010 Speaker Recognition Evaluation," in *2010 NIST Speaker Recognition Evaluation (SRE) Workshop*, Brno, Czech Republic, 24-25 June 2010.

[24] N. Brummer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing," Tech. Rep., 2011. [Online]. Available: https://sites.google.com/site/nikobrummer

[25] *The NIST Year 2010 Speaker Recognition Evaluation Plan*. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html

[26] *The NIST Year 2012 Speaker Recognition Evaluation Plan*. [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf