Combination of Cepstral and Phonetically Discriminative Features for Speaker Verification

Achintya Sarkar, Cong-Thanh Do, Viet-Bac Le, Claude Barras

Combination of Cepstral and Phonetically Discriminative Features for Speaker Verification

Achintya K. Sarkar, Cong-Thanh Do, Viet-Bac Le and Claude Barras, *Member, IEEE*

***Abstract*—Most speaker recognition systems rely on short-term acoustic cepstral features for extracting the speaker-relevant information from the signal. But phonetic discriminant features, extracted by a bottle-neck multi-layer perceptron (MLP) on longer stretches of time, can provide a complementary information and have been adopted in speech transcription systems. We compare the speaker verification performance using cepstral features, discriminant features, and a concatenation of both followed by a dimension reduction. We consider two speaker recognition systems, one based on maximum likelihood linear regression (MLLR) super-vectors and the other on a state-of-the-art i-vector system with two session variability compensation schemes. Experiments are reported on a standard configuration of NIST SRE 2008 and 2010 databases. The results show that the phonetically discriminative MLP features retain speaker-specific information which is complementary to the short-term cepstral features. The performance improvement is obtained with both score domain and feature domain fusion and the speaker verification equal error rate (EER) is reduced up to 50% relative, compared to the best i-vector system using only cepstral features.**

***Index Terms*—Speaker verification, i-vector, multi-layer perceptron, bottleneck features, PCA, LDA, PLDA**

I. Introduction

ACOUSTIC cepstral features, extracted from short-term speech frames of 20-30 ms, are widely used in state-of-the-art speaker verification systems [1]. Since a few years, discriminative features, as extracted by a multi-layer perceptron (MLP), have been adopted in automatic speech recognition (ASR) systems in combination with short-term cepstral features thanks to their relevance and effectiveness [2], [3], [4]. The extraction of MLP feature makes use of temporal information which spans much longer stretches of time (typically 300-500 ms), compared to the extraction of cepstral features. MLP features used for ASR may consist of phoneme posterior probabilities (Tandem connectionist features [5], [6]) or the linear outputs of the neurons in the bottle-neck layer of the MLP. The latter ones, known as bottle-neck features, have been found to be more suitable in the framework of hidden markov model (HMM)-gaussian mixture model (GMM) based ASR [7]. Both probabilistic and bottle-neck features contain a phonetic information which is derived by the MLP from long-term speech frames. This longer stretch of time ensures that a significant phonetic information from speech signal is taken into account in the calculation of each MLP feature vector. Such features may also keep timbre-specific information and thus be relevant for a speaker recognition system. Alternatively, the MLP can also be trained to compute the target speakers posterior probabilities, using either the output layer [8] or the bottle-neck layer [9], [10] as features. Stoll et al. compared speaker- and phonetic-discriminative features for a speaker recognition task and got slightly better performance with the phonetic-based features [11]; however, simply concatenating MLP and cepstral features did not help in improving the speaker recognition performance. In a previous work [12], we also observed that augmented features, consisting of phonetically discriminative MLP and cepstral features, do not outperform the cepstral features. We have thus proposed to reduce the dimension of the augmented features, using principal component analysis (PCA), which helps in improving the speaker verification performance compared to the performance obtained with cepstral features [12]. However, these results were obtained with a baseline GMM-universal background model (UBM) system [13] and they need to be confirmed in a more performing framework.

In this paper, we study the effectiveness of combining cepstral features with phonetically discriminative features in a state-of-the-art speaker verification system with session variability compensation technique and we investigate linear discriminant analysis (LDA) on augmented features to discriminate the speakers. We consider two speaker verification systems, one is based on the state-of-the-art i-vector approach [14] and the other on maximum likelihood linear regression (MLLR) super-vectors [29], [15]. For session variability compensation, we explore two recently developed techniques namely eigen factor radial (EFR) [16] and probabilistic LDA (PLDA) [17]. We show that augmented features improve the speaker verification performance in contrast to several previous studies [11], [12]. The system performances are demonstrated on a standard task of NIST speaker recognition evaluation (SRE) 2008 and 2010 core condition.

II. Feature extraction

A. Cepstral features

Cepstral feature are estimated on the telephone bandwidth (0-4kHz) every 10ms, using a 30 ms analysis window. For each frame the cubic root of the Mel scale power spectrum is computed, followed by an inverse Fourier transform, and 12 LPC-based cepstral coefficients are extracted, using a process similar to that of perceptual linear predictive (PLP) coefficients

A. K. Sarkar, C.-T. Do and C. Barras are with LIMSI-CNRS, Université Paris-Sud, B.P. 133, 91403 Orsay Cedex, France. Viet-Bac Le is with Vocapia Research, 28 rue Jean Rostand, Parc Orsay Université, 91400 Orsay, France. E-mails: {sarkar, ctdo, barras}@limsi.fr; levb@vocapia.com. This work was partly realized through the QUAERO Program and the QCOMPERE project, funded by OSEO (French State agency for innovation) and ANR (French national research agency), respectively.

[18]. Cepstral mean removal and variance normalization are carried on independently for each speaker utterances. The 39-dimensional acoustic feature vector consists of 12 cepstral coefficients and the log energy, along with the first and second derivative coefficients computed over a window of 5 and 7 frames, respectively. Speech fundamental frequency F_0, which reflects the vocal fold vibration rate, can also be useful for speaker verification and complement the spectral envelope [19], [20], [21]. In this respect, a 3-dimensional pitch feature vector (pitch, Δ and $\Delta\Delta$ pitch) is extracted, using the autocorrelation method [22] coupled with linear interpolation in order to avoid zero values in the unvoiced segments. The pitch feature vector is added to the original PLP features, resulting in a 42-dimensional cepstral feature vector (PLP+F_0). These features are used as the baseline cepstral features and, henceforth, will be abbreviated as PLP.

B. Discriminative features

The MLP features are generated in two steps. The first step is raw features extraction which constitutes the input layer to the MLP neural network. In this work, the TRAP-DCT (TempoRAl Pattern -Discrete Cosine Transform) [7] is used as raw features. The TRAP-DCT features are obtained from a 19-band Mel scale spectrogram, using a 30 ms window and a 10 ms offset, similar to [23] on broadcast data. A discrete cosine transform (DCT) is applied to 500 ms window of each band from which 25 first DCT coefficients are retained. The retained DCT coefficients are then concatenated together. In total, the raw features have, thus, $19 \times 25 = 475$ DCT coefficients. The raw features are then input to a 4-layer MLP [3] with the bottle-neck architecture [7]. The size of the third layer (the bottle-neck) is equal to the desired number of features (39).

In a second step, the raw features are processed by the MLP and the features are not taken from the output layer of the MLP but from the hidden bottle-neck layer and de-correlated by a PCA whitening transformation. No speaker normalization and adaptation technique was applied on the raw features like VTLN or SAT/CMLLR [24] or on the MLP features like HLDA, phonetic MLLR adaptation [25], [24]. These normalization techniques may improve the ASR performance but remove a more speaker specific information. The MLP feature vector has finally 39 dimensions. An illustration of MLP (bottle-neck) feature extraction is shown in Fig. 1.

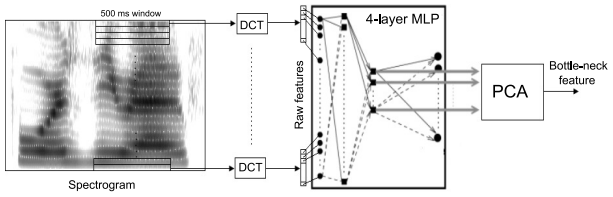

Fig. 1. MLP (bottle-neck) features extraction using a 4-layer MLP neural network. The input features are TRAP-DCT, extracted from 500 ms windows in the sub-bands of short-term spectrogram [3], [7]. PCA is applied to de-correlate the 39-dimensional feature vector taken from the bottle-neck layer.

The MLP neural network is trained using ICSI Quicknet software [26] on about 2000 hours of conversational telephone speech (CTS) data, mainly from the Switchboard, CallHome or Fisher databases provided by the LDC [27]. The phonetic segmentation was obtained through forced alignment. Since the amount of data for training the MLP is very large, efficient training procedure should be implemented. In our work, a simplified training scheme, proposed in [6], was applied for the training. Following this scheme, the conversation sides are randomized and split in three non-overlapping subsets, used in 6 training epochs with fixed learning rates. The first three epochs use only 13% of data, the next two use 26%, the last epoch uses 52% of the data, with the remainder used for cross-validation to monitor the performance. The MLP has 138 targets, corresponding to the individual states for each phone and one state for the additional pseudo phones (silence, breath, filler-word).

C. Feature Dimension Reduction

Two techniques are considered for reducing the dimension of the augmented features resulting from the concatenation of MLP (39 dimension) and PLP (42 dimension), namely principal component analysis (PCA) and linear discriminant analysis (LDA) [28]. With PCA, the projection space is generated through eigen value decomposition of the covariance matrix estimated using augmented features pooled over many non-target speakers. With LDA, the transformation matrix aims to maximize the ratio of between-class scatter S_B and within-class scatter S_W.

III. Speaker verification systems

We consider two approaches to evaluate the speaker verification performance of the proposed features; one is based on maximum likelihood linear regression (MLLR) super-vectors and the other relies on the standard i-vector approach.

A. MLLR super-vector

In the MLLR super-vector system [15], [29], speakers or utterances are represented by a super-vector formed by row-wise stacking the element of the respective speaker or utterance MLLR transformation [30]. The MLLR transformation for a speaker is estimated with respect to a universal background model (UBM) in the maximum likelihood (ML) sense using his/her training data, without any speech transcription, as

$$\hat{\mu}_k = A\mu_k + b, \quad \hat{\Sigma}_k = \Sigma_k \tag{1}$$

where μ_k and Σ_k represent the mean and co-variance matrix of the kth Gaussian of the UBM model and $\hat{\mu}_k$ and $\hat{\Sigma}_k$ are the adapted model parameters. The same MLLR transformation (A, b) is shared by all Gaussians. Then, the MLLR transformation matrix A of the particular speaker is stacked row wise to form the representative MLLR super-vector. The bias b did not provide measurable gains in our experiments and is not further considered. It results in a $F \times F$ dimensional super-vector depending on the dimension F of the feature vectors.

B. i-vector

The i-vector system characterizes speakers and utterances with vectors obtained by projecting their speech data onto a *total variability space* T where speaker and channel information is dense [14]. It is generally expressed as:

$$S = m + Tw \tag{2}$$

where w is called an *i-vector* and m and S are the GMM super-vector of the speaker independent UBM and speaker adapted model, respectively. It was implemented using the Bob toolkit [31], [32].

IV. Session variablity compensation & scoring

During the test phase, the i-vector or the MLLR super-vector of the test utterance is scored against the claimant speaker specific vector obtained in the training phase, after a post-processing of the vectors for session variability compensation. We consider two techniques most commonly used in the i-vector framework.

1) Eigen Factor Radial (EFR): The i-vector w is iteratively length normalized to compensate the session variability as per [16]. During test, the score between the length normalized i-vector of claimant $\hat{w}_{cl}$ and test utterance $\hat{w}_{tst}$ is calculated through a Mahalanobis distance normalized with the within-class covariance matrix computed using data pooled from many non-target speakers. In the case of MLLR based system, high-dimensional MLLR super-vectors are first projected onto a LDA space in order to reduce the dimension and better discriminate the speakers. Afterward, LDA projected MLLR super-vectors are length normalized and scored similarly to the i-vectors.

2) Probabilistic LDA (PLDA): PLDA is a generative modeling technique which decomposes the i-vector into several components as:

$$w = \mu_w + \Phi y_s + \Gamma z + \epsilon \tag{3}$$

where Φ and Γ are rectangular matrices representing the *eigen voice* and *eigen channel* subspace respectively. y_s and z are called the speaker and channel factor, respectively, with a priori normal distribution. ϵ indicates the residual noise. In test phase, the score between the i-vector of claimant w_{cl} and test utterance w_{tst} is calculated as:

$$score(w_{cl}, w_{tst}) = \log \frac{p(w_{cl}, w_{tst}|\theta_{tar})}{p(w_{cl}, w_{tst}|\theta_{non})} \tag{4}$$

with hypothesis θ_{tar} that w_{cl} and w_{tst} are from the same speaker and hypothesis θ_{non} that they are from different speakers. For details see [17]. MLLR super-vectors are processed similarly than i-vectors without any prior LDA.

V. Experimental setup

Experiments are preformed on male speakers of two standard tasks of NIST SRE 2008 (task 7, tel-tel) and 2010 (task 5, tel-tel) as per NIST evaluation plans [33], [34]. There are 1270 and 5200 utterances, respectively for NIST 2008 and 2010 for training 1270 and 5200 target models. All utterances are 5 minutes long with around 2.5 minutes speech duration. For the experiments on NIST SRE 2008, the *total variability space* T is trained using 12399 non-target speech utterances collected from various database (NIST 2004-05, Switchboard II part 1, 2 & 3; Switchboard cellular part 1 & 2, about 15 sessions per speaker; 890 speakers). This dataset is also used for implementing PCA, LDA, EFR and PLDA in both MLLR super-vector and i-vector systems. The reference speaker label is used for training the LDA and PLDA projections. In the case of PCA and LDA in augmented feature domain (i.e. concatenation of MLP and PLP), the file-wise mean vector is considered. For PLDA, both speaker and channel factors are varied to find the best speaker verification performance (with a step of 50 upto the dimension of the vector). MLLR adaptation is performed using a single iteration. For SRE 2010 experiments, 6947 additional utterances are taken from SRE 2006 and 2008 for training the T space, EFR and PLDA. However, the LDA or PCA projection matrices used are the ones estimated on SRE 2008 development set. The dimension of the i-vector is 400 for both SRE 2008 and 2010 systems. The UBM consisting of 512 Gaussians with diagonal covariance matrices is trained using non-target data from NIST SRE 2004. In the case of MLP+PLP augmented features followed by LDA or PCA, a dedicated i-vector or MLLR system is implemented on the projected features. However, UBM size, UBM training data, i-vector dimension, number of iterations for total variability space, PLDA training and procedure were fixed on SRE 2008 development set and kept identical for the experiments on SRE 2010 test set. The system performance is measured using the equal error rate (EER).

VI. Results and Discussion

For analysis, speaker verification system performances are presented with EFR and PLDA session variability technique on task 7 (tel-tel) of NIST SRE 2008 core condition. The best system is selected according to the lowest EER. Then, system performances for the best configuration are given on SRE 2010.

A. Performance on SRE 2008 development set

In this section, we compare the performance of a speaker verification system with or without augmented feature on task 7 of NIST SRE 2008 core condition. The optimal PCA or LDA projection size is selected based on the lowest EER for different values of projection as shown in Fig. 2 based on the respective system with EFR.

From Table I, it can be observed that a system using a simple concatenation of the MLP+PLP features without any further projection fails to improve upon the baseline MLLR system.

TABLE I
Comparison of speaker verification performance with or without PCA/LDA on augmented feature system with the baseline systems on task 7 of NIST SRE 2008 core condition for different configurations.

System	Features/dim.	Opt. proj.	%EER	
			EFR	**PLDA**
MLLR systems				
Baseline systems	MLP/39	-	5.08	3.97
	PLP/42	-	4.23	4.43
Augmented features	MLP+PLP/81	-	7.01	6.02
	MLP+PLP/81	PCA40	**3.44**	**2.65**
	MLP+PLP/81	LDA45	**3.39**	**3.20**
Score fusion	MLP/39 & PLP/42		**3.59**	**2.92**
i-vector systems				
Baseline systems	MLP/39	-	3.41	2.61
	PLP/42	-	2.51	2.05
Augmented features	MLP+PLP/81	-	**2.01**	**1.80**
	MLP+PLP/81	PCA70	**1.85**	**1.58**
	MLP+PLP/81	LDA70	**1.84**	**1.63**
Score fusion	MLP/39 & PLP/42		**2.48**	**1.61**

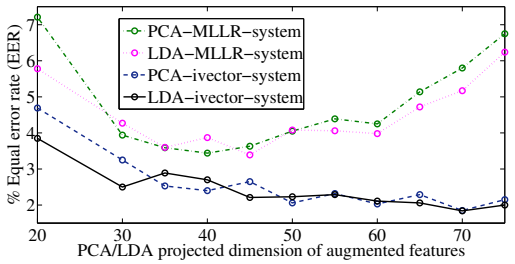

Fig. 2. Speaker verification performance of MLLR super-vector and i-vector systems with EFR for different PCA/LDA projected dimension of augmented features (MLP+PLP) on task 7 (tel-tel) of NIST SRE 2008 core condition.

Conversely, when projected either with PCA or LDA, the system using augmented features shows a remarkably lower EER compared to the best system with standalone features for both MLLR and i-vector systems with the different session variability and scoring techniques, a relative improvement above 20% for the MLLR systems and i-vector systems. The late fusion of standalone MLP and PLP feature based systems in the score domain also provides a comparable reduction of the EER. Thus, both MLP and PLP contain complementary speaker related information when integrated into a speaker verification system using a current session variability compensation technique in contrast to [11], [12]. The i-vector system yields a better performance than the MLLR based system for both EFR and PLDA. In the case of EFR, augmented features projected with LDA show a slightly better performance than with PCA. Conversely, for PLDA, a slightly better performance is observed with PCA, which could be due to a complementarity between PCA and PLDA.

B. Performance on NIST 2010 SRE

In this section, we further present the speaker verification performance on task 5 of NIST SRE 2010 core condition for the i-vector system only, using the PLDA parameters which where found optimal on NIST SRE 2008.

From Table II, we can observe a similar pattern than on NIST SRE 2008. The combination of MLP with PLP features result in a remarkable improvement of the speaker verification performance compared to the systems which uses standalone features for EFR or PLDA session variability compensation schemes. Departing from the observation on the development set, LDA only slightly improves the result compared to the raw concatenated features, while PCA actually degrades the performances. When using EFR scoring, feature fusion results in a better system than score fusion. Compared to the best performing cepstral-based i-vector system with PLDA scoring at 2.25% EER, the concatenation with MLP features followed by a LDA projection results in a 50% relative improvement, at 1.13% EER, and score fusion behaves similarly, halving the EER to 1.12%. More generally, the detection error trade-off (DET) curves presented on Fig. 3 for the PLDA scoring case show that the LDA projection provides the best performance for a large range of operating points.

TABLE II
Seaker verification performance on task 5 of NIST SRE 2010 core condition with i-vector *for different session variability and scoring techniques.*

System	Features/dim.	Opt. proj.	%EER EFR	%EER PLDA
Baseline systems	MLP/39	-	2.33	2.01
	PLP/42	-	2.55	2.25
Augmented features	MLP+PLP/81	-	**1.48**	**1.15**
	MLP+PLP/81	PCA70	**1.69**	**1.42**
	MLP+PLP/81	LDA70	**1.40**	**1.13**
Score fusion	MLP/39 & PLP/42		**1.97**	**1.12**

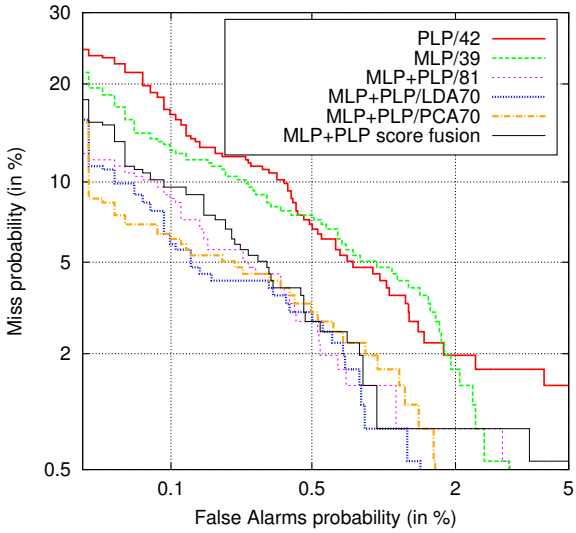

Fig. 3. DET curves corresponding to the PLDA systems presented in Table II.

VII. Conclusion

Recently, phonetically discriminative features extracted from long-term temporal windows with a bottle-neck MLP were found to be complementary to the cepstral features for ASR systems. In this work, we explored the combination of PLP cepstral features and MLP features in the context of speaker verification on two standard tasks of NIST SRE 2008 and 2010 core condition. We observed that a system using concatenated features remarkably outperforms the standalone systems, in a state-of-the-art i-vector framework with EFR or PLDA session variability compensation and scoring. It generally helps to project the augmented features onto a lower-dimensional space using PCA or LDA, however the gains obtained with PCA on the development set were not observed on the test set. Using LDA-projected concatenated features, the speaker verification equal error rate was reduced by about 50% relative compared to the best cepstral i-vector system on SRE 2010. Late fusion in score domain of the MLP and PLP systems also provided a similar improvement compared to the corresponding standalone systems, and even slightly outperformed the feature-domain fusion in the best configuration of i-vector systems with PLDA scoring.

These results confirm, as was observed in previous study [12], that the phonetically discriminative MLP features retain speaker-specific information which is complementary to the short-term cepstral features. Furthermore, their combination is effective both in score and feature domain and provides an important gain in the context of a state-of-the-art speaker verification system.

REFERENCES

[1] Kinnunen, T. and Li, H., "An overview of text-independent speaker recognition: from features to supervectors", *Speech Communication*, 52(1):12–40, Jan. 2010.

[2] Morgan, N., et al., "Pushing the envelope - Aside", *IEEE Signal Processing Magazine*, 22(5):81–88, Sep. 2005.

[3] Fousek, P., Lamel, L. and Gauvain, J.-L., "Transcribing broadcast data using MLP features", *INTERSPEECH*, pp. 1433–1436, September 22-26, Brisbane, Australia, 2008.

[4] Valente, F., Magimai-Doss, M., Plahl, C., Ravuri, S. and Wang, W., "A comparative large scale study of MLP features for Mandarin ASR", *INTERSPEECH*, pp. 2630–2633, September 26-30, Makuhari, Japan, 2010.

[5] Hermansky, H., Ellis, D., and Sharma, S., "Tandem connectionist feature stream extraction for conventional HMM systems," *ICASSP*, vol. III, pp. 16351638, Istanbul, Turkey, June, 2000.

[6] Zhu, Q., Stolcke, A., Chen, B.Y. and Morgan, N., "Using MLP features in SRI's conversational speech recognition system", *INTERSPEECH*, pp. 2141–2144, September 04-08, Lisbon, Portugal, 2005.

[7] Grezl, F. and Fousek, P., "Optimizing bottle-neck features for LVCSR", *IEEE ICASSP*, pp. 4729–4732, March 30 - April 04, Las Vegas, USA, 2008.

[8] Heck, L. P., Konig, Y., Sonmez, M. K. and Weintraub, M., "Robustness to telephone handset distortion in speaker recognition by discriminative feature design", *Speech Communication*, 31(2-3):181–192, Jun. 2000.

[9] Wu, D., Morris, A. and Koreman, J., "MLP internal representation as discriminative features for improved speaker recognition", *NOLISP'05*, pp. 72–80, April 19-22, Barcelona, Spain, 2005.

[10] Yaman, S., Pelecanos, J. and Sarikaya, R., "Bottleneck features for speaker recognition", *Odyssey'12*, pp. 105–108, June 25-28, Singapore, 2012.

[11] Stoll, L., Frankel, J. and Mirghafori, N., "Speaker recognition via nonlinear discriminant features", *NOLISP'07*, pp. 27–30, May 22-25, Paris, France, 2007.

[12] Do, C.-T., Barras, C., Le, V.-B. and Sarkar, A.K., "Augmenting short-term cepstral features with long-term discriminative features for speaker verification of telephone data", *INTERSPEECH'13*, August 25-29, Lyon, France, 2013.

[13] Reynolds, D., Quatieri, T. and Dunn, R., "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, 87:19–41, 2000.

[14] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P., "Front-End Factor Analysis for Speaker Verification", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19,pp. 788-798, 2011.

[15] Sarkar, A., Bonastre, J.-F. and Matrouf, D., "Speaker verification using m-vector extracted from MLLR super-vector", *EUSIPCO*, pp. 21–25, August 27-31, Bucharest, Romania, 2012.

[16] Bousquet, P. M., Matrouf, D. and Bonastre, J. F., "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition," *Proc. of INTERSPEECH*, 2011.

[17] Prince, S., "Computer Vision: Models Learning and Inference," *Cambridge University Press*, 2012.

[18] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, 87(4):1738–1752, 1990.

[19] Adami, A. G., Mihaescu, R., Reynolds, D. A. and Godfrey, J. J., "Modeling Prosodic Dynamics for Speaker Recognition," *Proceedings of ICASSP*, pp. IV-788-91, 2003.

[20] Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D. and Xiang, B., "The superSID project: exploiting high-level information for high-accuracy speaker recognition," *Proceedings of ICASSP*, pp. 784787, 2003.

[21] Shriberg, E., "High-level features in speaker recognition," *Lecture Notes in Artificial Intelligence*, Speaker Classification (C. Mueller Eds.), Springer, Heidelberg, Germany, vol. 4343, 2007.

[22] Boersma, P., "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proc. of the Institute of Phonetic Sciences*, vol. 17, pp. 97–110, University of Amsterdam, 1993.

[23] Le, V.B., Lamel, L. and Gauvain, J-.L., "Multi-Style MLP Features for BN Transcription," *Proceedings of ICASSP*, pp. 4866–4869, Dallas, TX, March 2010

[24] Tüske, Z., Plahl, C., Schlter, R., "A study on speaker normalized MLP features in LVCSR," *INTERSPEECH*, 1089-1092, 2011.

[25] Zhu, Q. , Chen, B. Y. , Morgan, N. and Stolcke, A., "On using MLP features in LVCSR.," *INTERSPEECH*, 2004.

[26] Johnson, D, Ellis, D, Oei, C., Wooters, C. and Faerber, P, "ICSI Quick-Net software package," http://www1.icsi.berkeley.edu/Speech/qn.html.

[27] Prasad, R., Matsoukas, S., Kao, C.-L., Ma, J. Z., Xu, D. X., Colthurst, T., Kimball, O., Schwartz, R. M., Gauvain, J.-L. and Lamel, L., "The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system," *INTERSPEECH*, pp. 16451648, 2005.

[28] Fukunaga, K., "Introduction to Statistical Pattern Recognition", *Academic Press*, 2nd Eds., pp. 445, 1990.

[29] Stolcke, A., Ferrer, L., Kajarekar, S. and Venkataraman, A., "MLLR transforms as features in speaker recognition", *INTERSPEECH*, pp. 2425–2428, September 04-08, Lisbon, Portugal, 2005.

[30] Leggetter, C. and Woodland, P., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, 9:171–186, 1995.

[31] Anjos, A., El Shafey, L., Wallace, R., Günther, M., McCool, C., Marcel, S., "Bob: a free signal processing and machine learning toolbox for researchers", *20th ACM Conference on Multimedia Systems (ACMMM)*, Nara, Japan, Oct. 2012.

[32] Khoury, E., El Shafey, L. and Marcel, S., "SPEAR: An open source toolbox for speaker recognition based on Bob," *Proc. ICASSP*, 2014.

[33] The NIST Year 2008 Speaker Recognition Evaluation Plan., "http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf"

[34] The NIST Year 2010 Speaker Recognition Evaluation Plan., "http://www.itl.nist.gov/iad/mig//tests/sre/2010/NIST_SRE10_evalplan.r6.pdf"