Augmenting Short-term Cepstral Features with Long-term Discriminative Features for Speaker Verification of Telephone Data

Cong-Thanh Do[1], *Claude Barras*[1], *Viet-Bac Le*[2], *Achintya K. Sarkar*[1]

[1]LIMSI-CNRS, Université Paris-Sud, 91403 Orsay Cedex, France
[2]Vocapia Research, Parc Orsay Université, 91400 Orsay Cedex, France
{ctdo, barras, sarkar}@limsi.fr, levb@vocapia.com

Abstract

Short-term cepstral features have long been chosen as standard features for speaker recognition thanks to their relevance and effectiveness. In contrast, discriminative features, calculated by a multi-layer perceptron (MLP) from much longer stretches of time, have been gradually adopted in automatic speech recognition (ASR). It has been shown that augmenting short-term cepstral features with long-term MLP (multi-layer perceptron) features makes it possible to improve significantly the performance of ASR. In this work, we investigate the possibility of augmenting short-term cepstral features with MLP features in order to improve the performance of text-independent speaker verification. We show, that, even though augmenting cepstral features with MLP features does not directly improve speaker verification performance, reducing the dimension of the augmented features, using principal component analysis (PCA), makes it possible to reduce, relatively, around 12% of the equal error rate (EER). Experiments are performed on telephone data of the 2008 NIST SRE (speaker recognition evaluation) database.

Index Terms: Speaker verification, multi-layer perceptron (MLP), principal component analysis (PCA), NIST SRE 2008, GMM-UBM

1. Introduction

Automatic speaker verification aims at verifying whether a given speech segment has been spoken by a claimed target speaker or not. Amongst the acoustic features used in state-of-the-art speaker verification systems, cepstral features, extracted from short-term speech frames of 20-30 ms, have been widely used [1]. Extracting cepstral features from short-term speech frames is relevant to the speaker modeling framework in which cepstral coefficients are assumed to be stationary random variables within a speech frame. On the other hand, dynamic features [2], which computed the time differences between the adjacent cepstral vectors, have usually been appended to the cepstral vectors.

Discriminative features, extracted by a trained multi-layer perceptron (MLP) $\mathbb{M}$, have been introduced [3] and gradually adopted in automatic speech recognition (ASR) systems thanks to their relevance and effectiveness [4, 5]. The extraction of MLP features makes use of temporal information which spans much longer stretches of time (from 0.5 to 1.0 second), compared to the extraction of cepstral features. In fact, augmenting short-term cepstral features with long-term MLP features makes it possible to improve significantly the performance of ASR.

This work has been partially financed by OSEO, the French State Agency for Innovation, under the Quaero program and the ANR project QCOMPERE.

MLP features, designed for ASR, could consist of phoneme posterior probabilities or the linear outputs of the neurons in the bottle-neck layer of the MLP $\mathbb{M}$. The latter one, known as bottle-neck features, has been found to be more suitable for classification application, namely ASR [4, 6]. Indeed, both probabilistic and bottle-neck MLP features contain phonetic information which is derived, by the MLP $\mathbb{M}$, from long-term speech frames covering certain numbers of phonemes. 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.

In text-independent speaker verification, the content spoken in the training and testing utterances could be completely different [7]. In fact, phonetic variability represents one adverse factor to accuracy in text-independent speaker recognition [1]. Therefore, the text-independent speaker verification system must take into account and handle the phonetic variability as well as the phonetic mismatch between train and test. In this respect, the phonetic information conveyed in the MLP features, designed for ASR, could be useful for text-independent speaker verification. In fact, the probabilistic or bottle-neck MLP features reflect the uncertainty of assessing the observation vectors to phonetic classes. Hence, these features could be useful in handling phonetic variability and phonetic mismatch between training and test, for text-independent speaker verification.

The MLP $\mathbb{M}$ can also be trained to compute the target speakers posterior probabilities, as in [8]. The resulting MLP features have been used separately with cepstral features and they were reported to outperform cepstral features in speaker recognition of telephone handset data. Following work of Wu et al. [9] reported similar improvement with this type of MLP-based features, on TIMIT database. Bottle-neck features have been also investigated in speaker recognition but these features do not outperform cepstral features [10]. In [11], Stoll et al. have investigated the augmentation of cepstral features with MLP-based features which consist of either phonemes or target speakers posterior probabilities. However, no significant gain has been reported on the system using the augmented features compared to the baseline system using cepstral features alone [11].

In this paper, we investigate the possibility of improving the performance of text-independent speaker verification by augmenting cepstral features with MLP features which are designed for ASR system at LIMSI [4]. Our work is different compared to the study of Stoll et al. [11] in two main points. First, instead of using MLP features consisting of target speakers or phonemes posterior probabilities, we use bottle-neck features to augment cepstral features. These bottle-neck features consist of the linear outputs of the neurons in the bottle-neck layer of the MLP, trained to discriminate phonetic classes [4]. Second, we

propose to use principal component analysis (PCA) to reduce the dimension of the augmented feature vector. The motivation for this dimension reduction using PCA is as follows. The bottle-neck MLP features might contain complementary phonetic information for cepstral features but they might contain also redundant information to the cepstral features. Therefore, using PCA to reduce the dimension of the augmented feature vector would help in maintaining complementary and reducing redundant information between MLP and cepstral features.

The paper is organized as follows. Section 2 presents the feature extraction, including cepstral and discriminative (MLP) features. After that, the reduction of the dimension of the augmented feature vector, using PCA, is introduced in section 3. Section 4 describes the experimental setup, including data and speaker verification architecture. The speaker verification results are introduced in section 5. Finally, section 6 concludes the paper.

2. Feature extraction

2.1. Cepstral features

Cepstral feature vector consists of 39 PLP-like (perceptual linear predictive) coefficients [12] derived from a Mel frequency spectrum estimated on the telephone bandwidth (0-8kHz), every 10ms. Cepstral mean removal and variance normalization are carried out on the basis of speech clusters, obtained after automatic speech segmentation and speaker clustering, resulting in a zero mean and unity variance for each cepstral coefficient. The 39-dimensional acoustic feature vector consists of 12 cepstral coefficients and the log energy, along with the first and second derivative coefficients.

Speech fundamental frequency F0 (or as perceived, pitch) reflects the vocal fold vibration rate. This is one of the most speaker-specific information from speech signal that can be useful for speaker verification [13]. In this respect, a 3-dimensional pitch feature vector (pitch, Δ and $\Delta\Delta$ pitch) is extracted, using autocorrelation method together with linear interpolation [4], and added to the original PLP features, resulting in a 42-dimensional cepstral feature vector (PLP+F0). These features are used as the baseline cepstral features.

2.2. Discriminative features

The MLP features are generated in two steps. The fisrt step is raw features extraction which constitutes the input layer to the MLP neural network $\mathbb{M}$. In this work, the TRAP-DCT (TempoRAl Pattern -Discrete Cosine Transform) [6] is used as raw features. The TRAP-DCT features are obtained from a 19-band Bark scale spectrogram, using a 30 ms window and a 10 ms offset. 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 $\mathbb{M}$ [14] with the bottle-neck architecture [6]. 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 $\mathbb{M}$ and the features are not taken from the output layer of the MLP $\mathbb{M}$ but from the hidden bottle-neck layer and decorrelated by a PCA transformation. The MLP feature vector has finally 39 dimensions. An illustration of MLP (bottle-neck) feature extraction is shown in Fig. 1.

The MLP neural network $\mathbb{M}$ is trained on about 2000 hours of conversational telephone speech (CTS) data which is similar to the CTS data used in [15]. Since the amount of data for training the MLP $\mathbb{M}$ is very large, efficient training procedure should be implemented. In our work, a simplified training scheme, proposed in [16], was applied for the training. Following this scheme, the training data 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 Quicknet[1] software was used to train the MLP. 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). The outputs of the MLP were normalized to range between 0 and 1 using the softmax function.

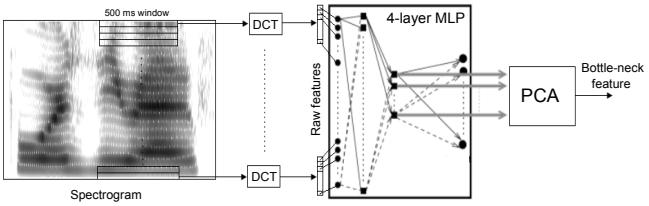

Figure 1: MLP (bottle-neck) features extraction using a 4-layer MLP neural network. The input features are TRAP-DCT, extracted from 500 ms windows in the subbands of short-term spectrogram [4, 6]. PCA is applied to decorrelate the 39-dimensional feature vector taken from the bottle-neck layer.

3. Feature dimension reduction using PCA

The cepstral features (C-dimensional) are augmented with the discriminative features (D-dimensional). The augmented feature vector $\mathbf{y}$ has cumulative dimension (L-dimensional, $L = C + D$) and could contain redundant information for speaker verification. We propose to reduce the dimension of the feature vector $\mathbf{y}$ using principal component analysis (PCA). To reduce the dimension of the augmented feature vector $\mathbf{y}$ by PCA, a transformation matrix $\mathbf{P}$ of $L \times L$ dimensions, whose columns are the principal components, is calculated. The augmented feature vector $\mathbf{y}$ are then linearly transformed to a lower dimension feature vector $\widehat{\mathbf{y}}$ using a matrix $\widehat{\mathbf{P}}$ of $L \times M$ dimensions ($M < L$), which contains M first principal components, following the equation:

$$\widehat{\mathbf{y}} = \widehat{\mathbf{P}}^T \mathbf{y}$$

where ^{T} denotes the transpose. The M-dimensional feature vector $\widehat{\mathbf{y}}$ is then used for training and testing of speaker verification system. To calculate the matrix $\mathbf{P}$, disjoint data, which is not used in train and test, is selected. Augmented feature vectors (L-dimensional) are extracted from this data and are put adjacently in a matrix $\mathbf{Y}$. After that, the matrix $\mathbf{P}$ is calculated from the data matrix $\mathbf{Y}$ by singular value decomposition (SVD) technique [17]. This matrix is calculated once and is the only matrix using for features projection.

4. Experimental setup

We made use of basic GMM-UBM (Gaussian mixture model - universal background model) [18] speaker verification system for evaluating the effectiveness of the augmentation of cepstral features with discriminative features. The UBM was trained

[1] http://www1.icsi.berkeley.edu/Speech/qn.html

on 243 utterances spoken by male speakers, extracted from the 2004 NIST SRE (speaker recognition evaluation) database [19]. There are 1270 targets speakers whose models are GMMs obtained by maximum a posteriori (MAP) adapting [20] of the UBM, using training data taken from the 2008 NIST SRE database [21]. Each target model has one utterance for MAP adaptation. The MAP adaptation was performed with 3 iterations and the MAP relevant factor r was set to 10.

In the current work, we evaluate the speaker verification system on telephone data (the MLP neural network $\mathbb{M}$ was already trained with English conversational telephone speech (CTS)). In this respect, we make use of the $6th$ and $7th$ evaluation tasks of the 2008 NIST SRE common evaluation conditions in which telephone data are used in training as well as in testing. In fact, speaker verification of telephone data is one of the standard evaluation conditions since performing speaker verification over telephone network is an essential task. More specifically, in the $6th$ evaluation task (DET6), all the trials involve only telephone speech in training and test. In the $7th$ evaluation task (DET7), all the trials involve only English language telephone speech in training and test. There are 12491 trials in the $6th$ and 6615 trials in the $7th$ evaluation tasks. For each trial, the log-likelihood of the test segment given the target model is normalized with AT-norm (adaptive T-norm) [22], using the scores (excluding the 5 highest ones) obtained when scoring the test segment against all the target speaker models.

Three types of acoustic features were used in the experiments, including 42-dimensional cepstral features PLP+F0, 81-dimensional MLP+PLP+F0 augmented features and M-dimensional reduced features obtained from the augmented features by using PCA. We denote these reduced features as MLP+PLP+F0-PCA features. In this work, we have tried three values of $M = \{40, 50, 60\}$ in order to evaluate the different degrees of redundant reduction from the augmented features. Disjoint data, consisting of 12392 utterances, from NIST SRE 2004, 2005 and Swichboard databases were selected to calculate the 81×81 dimensions matrix $\mathbf{P}$ which consists of the principal components. The data matrix $\mathbf{Y}$, for calculating $\mathbf{P}$, consists of 12392 81-dimensional augmented feature vectors (MLP+PLP+F0). These feature vectors were randomly selected, one feature vector per utterance, from the previously mentioned 12392 utterances. The result obtained with MLP features was also studied.

Fig. 2 shows an analysis of the cumulative variance conveyed in the principal components of the projection matrix $\mathbf{P}$. The cumulative variance conveyed in all the principal components (81) is 100%. It can be observed that the first 40, 50 and 60 principal components contain 78.2%, 89.2% and 95.7% of the total variance, respectively. In fact, only 50 first principal components contain already nearly 90% of the total variance. Therefore, it is thinkable that reducing the dimension of the augmented feature vectors would be relevant since a reasonable number of principal components contain already a significant cumulative variance.

5. Speaker verification results

The speaker verification results, in terms of equal error rates (EERs), are presented in Table 1. It can be observed that PLP+F0 features outperform MLP features, in terms of EERs. The big gap between PLP+F0 and MLP features indicates that the MLP (bottle-neck) features, generated by a MLP neural network trained to discriminate phonetic classes, are not relevant to be used alone for speaker verification. Linear fusion is performed with the scores obtained with PLP+F0 and MLP features. The best combination weights were 0.9 and 0.1 for the scores obtained with PLP+F0 and MLP features, respectively. It can be observed that the EERs, calculated from the fused scores, are lower than the EERs of the baseline system.

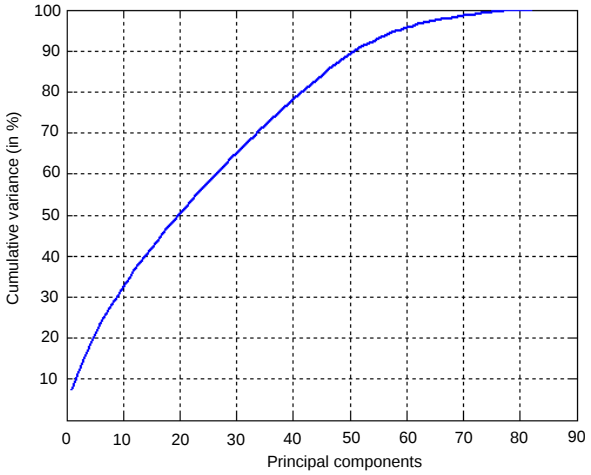

Figure 2: Cumulative variance conveyed in the principal components of the projection matrix $\mathbf{P}$. An almost linear increase in cumulative variance is observed between the first 20 and 50 principal components.

Table 1: Speaker verification results, in terms of equal error rates (EER, in %), obtained with cepstral (PLP+F0), discriminative (MLP), augmented (MLP+PLP+F0) and reduced (MLP+PLP+F0-PCA) features. Linear fusion of the scores, obtained with PLP+F0 and MLP features, are performed. The scores linear fusion (displayed) performs best with the combination weights equal 0.9 and 0.1 for PLP+F0 and MLP scores, respectively.

Features \ Evaluation tasks	DET6	DET7
PLP+F0 (baseline)	16.45	15.31
MLP	19.58	18.23
Score fusion (PLP+F0, MLP)	16.14	15.10
MLP+PLP+F0	16.61	15.95
MLP+PLP+F0-PCA ($M = 40$)	14.89	13.92
MLP+PLP+F0-PCA ($M = 50$)	**14.51**	**13.36**
MLP+PLP+F0-PCA ($M = 60$)	15.46	14.63

In contrast, the augmented features MLP+PLP+F0 do not help in improving the EER, compared to the baseline system. PCA has been applied to reduce the dimension of the MLP+PLP+F0 feature vector from 81 to 40, 50 and 60. The EERs obtained with the reduced features, of 40-, 50- and 60-dimensional, are lower than those of the baseline system, in both evaluation tasks (DET6 and DET7). It can be observed that the 50-dimensional reduced features (MLP+PLP+F0-PCA) give lower EERs compared to the 40- and 60-dimensional reduced features. In the DET6 evaluation task, the EER has been reduced 1.94%, absolutely, and 11.8%, relatively, with the system using 50-dimensional reduced features (MLP+PLP+F0-PCA), compared to the baseline system using cepstral PLP+F0 features. Similarly, in the DET7 evaluation task, there is 1.95%

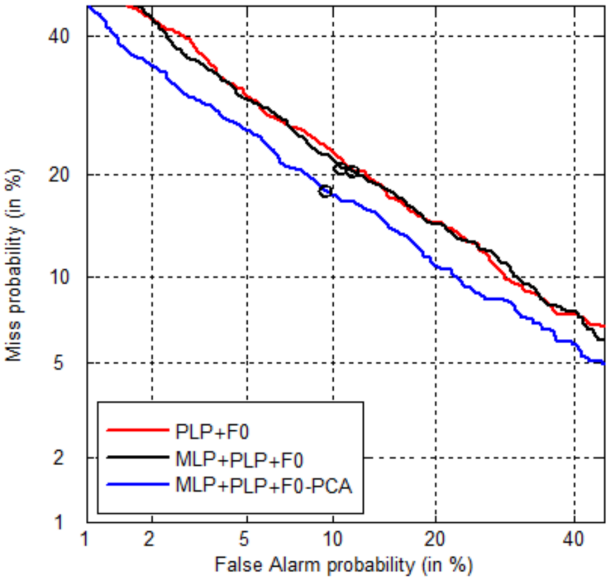

Figure 3: DET (detection error tradeoff) curves of speaker verification systems, corresponding with PLP+F0 (red), MLP+PLP+F0 (black) and MLP+PLP+F0-PCA (blue) features. The MLP+PLP+F0-PCA feature vectors have 50 dimensions ($M = 50$). All the trials involve only telephone speech in training and test (DET6).

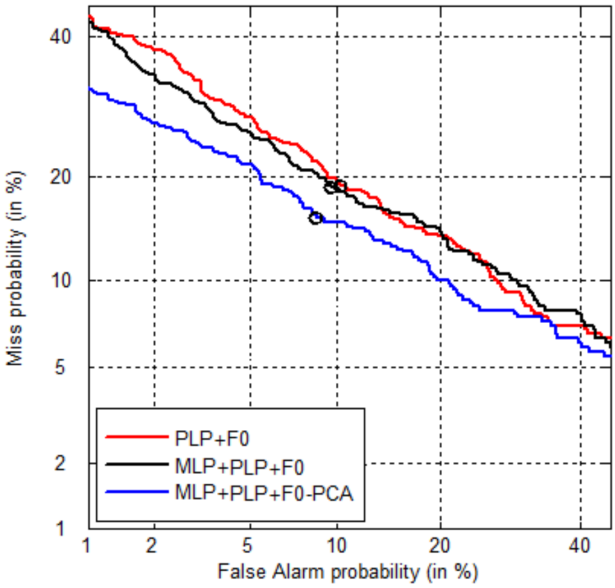

Figure 4: DET curves of speaker verification systems, corresponding with 3 types (cepstral, augmented and PCA-reduced) of features. The MLP+PLP+F0-PCA feature vectors have 50 dimensions ($M = 50$). All the trials involve only English language telephone speech in training and test (DET7).

absolute reduction and 12.74% relative reduction of the EER, compared to the baseline system, when the 50-dimensional reduced features are used.

The DET (detection error tradeoff) curves of the systems, using the PLP+F0, MLP+PLP+F0 and 50-dimensional MLP+PLP+F0-PCA features, are shown in Figs. 3 and 4. It can be observed that the DET curves of the system, using the 50-dimensional MLP+PLP+F0-PCA features, are entirely below the curves of the systems using PLP+F0 and MLP+PLP+F0 features, in both evaluation tasks. There is in contrast no clear difference between the DET curves of the systems using PLP+F0 and MLP+PLP+F0 features.

6. Conclusion

We have investigated the possibility of improving the performance of speaker verification, of telephone data, by augmenting short-term cepstral features (PLP+F0) with long-term discriminative MLP features. Experiments have shown that augmenting PLP+F0 features with MLP features does not help reducing the EER of speaker verification. Fusing the scores of the systems using PLP+F0 and MLP features helps in reducing the EER compared to the baseline system using PLP+F0 features. We have proposed to reduce the dimension of the augmented features (81-dimensional MLP+PLP+F0), using PCA, to reducing the redundant whereas keeping the complementary phonetic information between the MLP and PLP+F0 features. The reduced features, obtained with PCA, have given lower EERs compared to the baseline system using cepstral feature. The best EER reduction has been obtained with 50-dimensional MLP+PLP+F0-PCA features (around 12% relative reduction of the EER).

This study has shown that long-term information from speech signal, extracted from frequency subbands, is complementary for short-term cepstral features in speaker verification with telephone data. Such a related study using subband temporal information for speaker verification can be found in [23]. Future work will focus on the investigation of other efficient feature dimension reduction techniques, for instance LDA (linear discriminant analysis) [24], to reduce the dimension of the augmented MLP+PLP+F0 features.

7. 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] Furui, S., "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 34(1):52–59, Feb. 1986.

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

[4] Lamel, L., Gauvain, J.-L., Le, V.B., Oparin, I. and Meng, S., "Improved models for Mandarin speech-to-text transcription", *IEEE ICASSP*, pp. 4660–4663, May 22-25, Prague, Czech Republic, 2011.

[5] 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.

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

[7] Bimbot, F., et al., "A tutorial on text-independent speaker verification", *EURASIP Journal on Applied Signal Processing*, 24(4):430–451, 2004.

[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. 114–123, May 22-25, Paris, France, 2007.

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

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

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

[15] Prasad, R., et al., "The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system", *INTERSPEECH*, pp. 1645–1648, September 04-08, Lisbon, Portugal, 2005.

[16] 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.

[17] Jolliffe, I.T., "Principal component analysis", *Springer series in statistics*, Springer-Verlag, 2nd Eds., pp. 487, 2002.

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

[19] "The NIST year 2004 speaker recognition evaluation plan", http://www.itl.nist.gov/iad/mig/tests/spk/2004, 2004.

[20] Gauvain, J.-L. and Lee, C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, Apr. 1994.

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

[22] Sturim, D.E. and Reynolds, D.A., "Speaker adaptive cohort selection for T-norm in text-independent speaker verification", *IEEE ICASSP*, pp. 741–744, March 18-23, Philadelphia, USA, 2005.

[23] Do, C.-T. and Barras, C., "Cochlear implant-like processing of speech signal for speaker verification", *SAPA (Statistical and Perceptual Audition) Conference*, satellite workshop of INTERSPEECH, pp. 17–21, September 07-08, Portland, OR, USA, 2012.

[24] Jin, Q. and Waibel, A., "Application of LDA to speaker recognition", *ISCA ICSLP*, pp. 250–253, October 16-20, Beijing, China, 2000.