STUDY OF MUTUAL INFORMATION FOR SPEAKER RECOGNITION FEATURES

Guillermo Garcia, Thomas Eriksson

Department of Signals and Systems,
Chalmers University of Technology, 412 96 Göteborg, Sweden
phone: + (46) 31 772 18 21 fax: +(46) 31 772 17 48
em1guill@chalmers.se, thomase@chalmers.se

ABSTRACT

Feature extraction is an important stage in speaker recognition systems since the overall performance depends on the type of the extracted features. In the framework of speaker recognition, the extracted features are mainly based on transformations of the speech spectrum. In spite of the great variety of features extracted from the speech, the common empirical approach to select features is based on a complete performance evaluation of the system. In this paper, we propose an information theory approach to evaluate the information about the speaker identity contained on the speech features. The results show that this approach can help on a more efficient feature selection. We also present an alternative AM-FM based magnitude representation of the speech that attains better performance than the MFCCs. Moreover, we show that phase information features can perform well in speaker verification systems.

1. INTRODUCTION

Speaker recognition aims at finding a set of characteristics that best represents a specific speaker voice. Formally, speaker recognition is defined as the process of automatically recognizing who is speaking based on information provided by speech signals. Speaker recognition contains two main phases: training and classification. In the training phase, the extracted features are used to create speaker models. In the classification phase, the speaker models are used to classify new input utterances to the system. Speaker recognition systems can also be categorized depending on their tasks in Speaker Identification (SID) and Speaker Verification (SV) systems [1]. SID systems determine from a set of predefined models to which of them belongs an input test utterance. Conversely, SV systems are employed to validate whether the speaker is who he or she claims to be.

Feature extraction, also called *front-end*, is the first stage in speaker recognition systems. In this stage, speech samples are transformed to a sequence of numerical descriptors called features. The extracted features contain individual and essential characteristics of the speakers and are used to estimate the speaker model parameters. In the speaker recognition framework, the basic *front-end* extracts and transforms the magnitude spectrum of the speech signal. The most common features in speaker recognition are the Mel Frequency Cepstral Coefficients (MFCCs) based on the known variation of the human ear's critical bandwidths, with filter-banks that are spaced linear at low frequencies and logarithmic at high frequencies [2]. The extracted features are often complemented with their delta and double delta, which capture the dynamic information of the features or in combination with other features [3]. On the other hand, focusing on the phase spectrum features, the Modified Group Delay Features (MGDFs) have proven to be useful in the extraction of speaker information from the phase of speech signals. These features are based on the group delay function [4].

The AM-FM representation can also be used as an alternative method for feature parametrization. This technique was used specially in the context of modulated signals [5]. In the area of speech processing, this technique is mainly utilized for formant tracking, speech synthesis, speech and speaker recognition. The AM-FM modeling decomposes the speech signals into decorrelated bandpass channels. Each channel is characterized in terms of its envelope (instantaneous amplitude) and phase (instantaneous frequency). In this work, we propose a methodology to extract features from the AM-FM representation, called instantaneous Amplitude Modulated features (AMFs) and Instantaneous Frequency features (IFFs).

Feature selection is not a new subject, studies comparing between magnitude spectrum features can be found in [6]. The importance of feature selection lies in the correlation between the extracted features and the performance of the system.

Information theory has proven to be useful in analyzing and selecting features in speaker recognition and in other applications. Several papers address the use of *mutual information* (MI) between features as a selection criterion [7]. The main advantage of using MI as a performance measure is that an analytical solution can be found and maximized.

In spite of the great variety of features extracted from the speech, the common empirical approach to select features is based on a complete performance evaluation of the system. In this paper, we propose an information theoretical approach to evaluate the information about the speaker identity contained on the speech features. We study magnitude and phase information features and the results show that this approach can help on a more efficient feature selection.

2. FEATURE EXTRACTION

In this work, we will analyze MFCCs, MGDFs, AMFs and IFFs. The reason for choosing these features, is to compare phase information and magnitude information features from an information theoretical perspective. Moreover, we analyze the phase of speech signals, such that discriminative information about the speaker can be used in the recognition process.

2.1 Modified Group Delay Features (MGDFs)

The MGDFs are features based on the group delay function [4]. This function provides a way of estimating the formants of the speech directly, without involving the process of

 ISSN 2076-1465

unwrapping the speech signal. The formants are resonances of the vocal tract containing discriminative information about the speaker. The MGDFs are defined as

$$\tau(\omega) = \zeta(\omega) \left| \frac{X_R(\omega) Y_R(\omega) - X_I(\omega) Y_I(\omega)}{(S(\omega))^{2\gamma}} \right|^{\alpha}, \quad (1)$$

where $X_R(\omega)$, $X_I(\omega)$ are the real and imaginary parts of the Fourier transform (speech spectrum) of the original speech signal $x[n]$, respectively. Additionally, $Y_R(\omega)$ and $Y_I(\omega)$ are the real and imaginary parts of the Fourier transform of $nx[n]$, respectively. $S(\omega)$ represents a smoothed version of the spectrum, $\zeta(\omega)$ defines the sign function of the group delay profile and α and γ are tuning parameters with typical values of 0.1 and 0.9, respectively. After obtaining the MGDFs, the DCT is applied to decorrelate the feature set. Finally, the feature set is truncate to the desired number of coefficients.

2.2 The AM-FM Parametrization

This type of representation decomposes the speech signal into decorrelated bandpass channels, each of them characterized by its envelope and phase. We will describe the steps for computing the instantaneous amplitude and frequency. Two steps are common for both feature extractors: the bandpass filtering of the speech signal and the computation of the analytical representation for the filtered signal. Then, separate steps are required to extract each of the features. We must denote that the speech signal should be pre-processed before the AM-FM parametrization, i.e., the speech signal should have been preemphasized and framed and each frame have to be windowed.

Bandpassing the Speech Signal: The first step is to bandpass the speech signal such that a number of decorrelated bandpass channels are attained. Three main requirements are specified for the design of the filters: central frequency, bandwidth and the type of the filter. The central frequencies are similar to the frequencies used in MFCCs, the bandwidth is defined by the perceptual critical band [8], and the filter type is a bandpass finite impulse response (FIR) filter.

Computation of the Analytical Signal: In order to characterize a single instantaneous frequency and amplitude for a real-valued signal, the analytical signal $x^{\mathrm{a}}(t)$ is constructed from each of the bandpass filtered signal as

$$x^{\mathrm{a}}(t) = x(t) + j\hat{x}(t), \quad (2)$$

where $\hat{x}(t)$ is the Hilbert transform of $x(t)$ [5]. The analytical signal approach provides a signal with no negative frequency components.

2.3 Amplitude Modulated Features (AMFs)

To extract the AMFs, the framed speech signal is bandpassed such that a waveform $v_i(t)$ is obtained for each i-th channel and for each frame. Then, the magnitude of the analytical representation can be computed for each speech frame as

$$a_i(t) = \sqrt{v_i^2(t) + \hat{v}_i^2(t)}, \quad (3)$$

where $\hat{v}_i(t)$ is the Hilbert transform of the i-th bandpassed waveform $v_i(t)$. Subsequently, we average the $a_i(t)$ over each frame and apply the logarithm in order to obtain a smooth representation of the magnitude and to emphasize the small variations of the signal for each channel.

$$A_i = \log E_{a_i}[a_i(t)], \quad (4)$$

$$= \log\left(\frac{1}{T}\sum_{t=1}^{T} a_i(t)\right), \quad (5)$$

where $E[\cdot]$ represents the expectation, and A_i are the average smoothed amplitude features for each i-th channel and T is the size of the frame. Finally, we decorrelate the A_i using again the DCT transformation.

2.4 Instantaneous Frequency Features (IFFs)

The IFFs are obtained from the analytical representation of the waveform after bandpass filtering the speech frame. The instantaneous frequency can be computed as the derivative of the instantaneous phase defined as

$$f_i(t) = \frac{1}{2\pi}\frac{\partial}{\partial t}\left[\arctan\left(\frac{\hat{v}_i(t)}{v_i(t)}\right)\right]. \quad (6)$$

An alternative method to compute the IFFs is to use the analytical form of the waveform, i.e., $v_i^{\mathrm{a}}(t) = v_i(t) + j\hat{v}_i(t)$. First, the instantaneous frequency can be estimated as [9]

$$f_i(t) = \frac{1}{4\pi}\arg\left[-\left(v_i^{\mathrm{a}}[t+1]\right)\left(v_i^{\mathrm{a}}[t-1]\right)^*\right], \quad (7)$$

where $v_i^{\mathrm{a}}[t-1]^*$ is the conjugate of the analytical representation of the waveform $v_i^{\mathrm{a}}[t-1]$. Then, $f_i(t)$ is averaged over each frame, in order to obtain a smooth representation of the instantaneous frequency for each filtered signal.

$$F_i = E_{f_i}[f_i(t)], \quad (8)$$

$$= \frac{1}{T}\sum_{t=1}^{T} f_i(t), \quad (9)$$

where T is the frame size. Finally the DCT-transform is computed and the new set of features are obtained.

3. SPEAKER MODELING

Gaussian Mixture Models (GMMs) have become the dominant approach for modeling speakers over the last years [10]. Given a feature vector x_t, the probability density function (pdf) of the speaker features $p(x_t|\lambda)$ can be approximated as a GMM

$$p(x_t|\lambda) = \sum_{k=1}^{K} \mathrm{w}_k \mathcal{N}(x_t|\mu_k, \mathbb{C}_k), \quad (10)$$

where $\mathcal{N}(x_t, \mu_k, \mathbb{C}_k)$ is a Gaussian distribution, μ_k is the mean vector, $\mathbb{C}_k$ is the covariance matrix and w_k is the weight of the k-th Gaussian distribution. The GMM can also be defined by a set of parameters, i.e., $\lambda = \{\mathrm{w}_k, \mu_k, \mathbb{C}_k\}_{k=1}^{K}$, estimated by the EM algorithm. In SV systems, two models are defined: the target model and the impostor model. The impostor model; also known as the Universal Background Model (UBM), is first trained using the Expectation Maximization (EM) algorithm and a pool of speakers different from the speaker we would verify. Then, the speaker model is derived from adapting the mean vectors of the UBM using *Maximum a Posteriori* (MAP) [11].

4. CLASSIFICATION

The extracted features from a speaker test utterance $\{x_t\}_{t=1}^{T}$ are compared against the speaker GMM. The speaker recognizer computes the log-likelihood of a given speaker model λ for the test utterances as

$$\mathcal{L}(x|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda). \tag{11}$$

Depending on the likelihood values obtained and the task, the system will emit a decision: the speaker is accepted/rejected or identified. To evaluate the performance of a speaker recognition system, we use the probability of error $P_e = \Pr[s \neq \hat{s}]$ where s represents the identity of the true speaker and $\hat{s}$ the decision emitted by the recognition system. Focusing on the speaker recognition tasks, SV is a statistical hypothesis test between two hypotheses [10]: the target and the impostor model. Moreover, each trial consists of a speaker test utterance and a claimed identity. From each trial, a log-likelihood ratio is computed and a score Θ is determined as

$$\Theta = \log\left(\frac{p(x|\lambda)}{p\left(x|\hat{\lambda}\right)}\right); \qquad \Theta \underset{\text{reject}}{\overset{\text{accept}}{\gtrless}} \tau \tag{12}$$

$$\Theta = \mathcal{L}(x|\lambda) - \mathcal{L}\left(x|\hat{\lambda}\right), \tag{13}$$

where λ denotes the hypothesis to accept an utterance $\{x_t\}_{t=1}^{T}$ as being produced by the target speaker. The impostor model or UBM ($\hat{\lambda}$) denotes the hypothesis to reject an utterance $\{x_t\}_{t=1}^{T}$ as being produced by the target speaker and τ is the threshold that minimizes the expected cost of errors.

5. INFORMATION THEORETICAL APPROACH

Our feature selection approach is based on the MI between the set of features and the speaker identity providing a way to quantify the amount of speaker information contained on each feature set. In [12], the bounds for the classification error tied to the mutual information between the feature set and the speaker identity are presented. The minimum possible classification error probability (P_e) for values less or equal to 0.5 is related to the MI between speakers and the feature vector sequences $\mathbf{C}$ as

$$I(\mathbf{C};S) \leq \log_2 M - 2P_e, \tag{14}$$

$$I(\mathbf{C};S) \geq \log_2 M - H(P_e) - P_e \log_2(M-1), \tag{15}$$

where M represents the total number of classes or speakers, which addresses the assumption of a uniform distribution of the classes S. Figure 1 presents an example of the bounds for $M = 32$ speakers for different kinds of features. Note that the bounds presented in (14) and (15) are based only on the extracted features and the number of speakers enrolled in the system. Further improvements on the bounds cannot be made without knowledge of the actual problem.

5.1 Analytical Expression for Mutual Information

To compute the MI, we let the set $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_Z\}$ contain all the feature vectors used in the classification. For simplicity, we will assume that independence between consecutive features so that the entropy is the sum of the frames entropies. Moreover, we will divide our database into M partitions, i.e., $\{\mathbf{c}_{s,t}\}_{t=1}^{T}$, $s = 1, \ldots, M$ with equal size T for each partition. Then, the MI can be defined as

$$I(\mathbf{c};S) = h(\mathbf{c}) - h(\mathbf{c}|S), \tag{16}$$

where $h(\mathbf{c})$ is the entropy of the features and $h(\mathbf{c}|S)$ is the conditional entropy of the features $\mathbf{c}$ averaged over all the speakers enrolled in the database M. The MI in (16) can be written as

$$I(\mathbf{c};S) = -E\left[\log_2 \sum_{s=1}^{M} \frac{1}{M} p(\mathbf{c}|s)\right] + E\left[\log_2 p(\mathbf{c}|S)\right], \tag{17}$$

where $p(\mathbf{c}|s)$ is the speaker pdf. We decided to use the GMMs since the speaker pdfs are already estimated with this method in speaker recognition systems. Substituting the expectations in (17) by the sample means and the pdfs by (10), we attain

$$I(\mathbf{c};S) = I(\mathbf{x};\lambda_S) \approx -\frac{1}{MT}\sum_{s=1}^{M}\sum_{t=1}^{T} \log_2 \sum_{i=1}^{M} \frac{1}{M} p(x_{s,t}|\lambda_i) + \frac{1}{MT}\sum_{s=1}^{M}\sum_{t=1}^{T} \log_2 p(x_{s,t}|\lambda_s). \tag{18}$$

where $\mathbf{x}$ is the feature vector and λ_i is the speaker model or class. The major advantage of using the MI as a performance measure instead of directly using the classification probability of error, is that the MI can be analytically computed and maximized more straightforward than the classification error. Using the MI estimator defined in (18), we can determine for any kind of proposed features which of them contains the largest MI with respect to the speaker and also the lowest P_e.

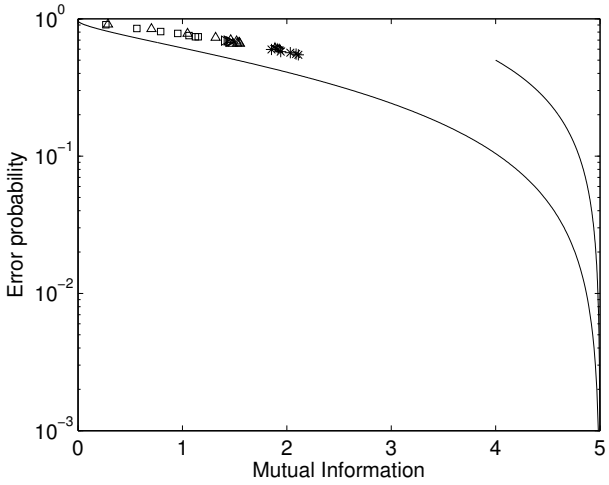

Figure 1: Bounds for the probability of error as a function of the mutual information $I(S;\mathbf{C})$ for a SID system consisting of $M = 32$ speakers. The squares, triangles and asterisks represent actual values from experiments performed with MGDFs, IFFs AMFs and MFCCs, (See section 6 for definitions) respectively.

6. EXPERIMENTAL SETUP

Two different experimental setups were determined. In the first setup, the experiments were conducted using the

"YOHO" database [13]. The first session of each of the first 30 speakers in the enrollment session has been used. In each speech file, the silences at the beginning and at the end were removed. Then, the speech file was segmented into frames of 25 ms and 10 ms of overlapping. Each frame was pre-emphasized and Hamming-windowed. Four different features sets were created: MGDFs, MFCCs, AMFs and IFFs. In the case of the MFCCs and the MGDFs, we used a 256-th Fast Fourier Transform (FFT) to extract the spectrum. Then, we compute the MGDFs and reduce their dimensionality to 22th-order after the DCT. The MFCCs were extracted with a 23th-order Mel triangular filter-bank. The first coefficient was discarded after the DCT, attaining 22th-order coefficients. For the AMFs and the IFFs, we use 23 different bandpass filters. The central frequency of the filters were similar to the MFCCs filters and the filters bandwidths were specified by the perceptual critical band [8]. The features extracted were used to train 32-mixtures GMM for each speaker. Then, we estimate the MI using the GMMs for each feature set. Moreover, we evaluate the system in open test conditions with 40 trials per speakers.

In the second setup, we verified that the selection of features for a larger database. The experiments were conducted using the female speakers from the 2004 NIST-SRE "core" corpus [14]. Each speech file consists of approximately five minutes one-side telephone conversion. A similar procedure to the first setup was used to create the frames. Then, MFCCs, MGDFs, AMFs and IFFs are obtained and warped with a 3 seconds Gaussian window [15]. Afterwards, deltas (Δ), double deltas ($\Delta\Delta$) and delta log-Energy ($\Delta \log E$) were computed. Later, a 512-mixtures UBM was trained and the speaker models were derived using MAP adaptation and their own feature set. Finally, we perform verification according to the 2004 NIST-SRE "core" corpus and compute the MI for the feature parameterizations with the best performance.

7. RESULTS

In figure 2, a comparison of the MI as a function of the number of coefficients for different features is presented. The results show that the AMFs contain the highest discriminative information about the speaker compared to the other feature sets. However, the results also show that the phase information features contain useful information about the speaker. Note that the MI is always an increasing function since the computation of the MI is done in *closed test* conditions (i.e., both model training and MI computations use the same database).

In figure 3, we present the MI and the P_e as a function of the number of coefficients for different features in *open test* conditions (i.e., different databases were used for model training and MI computation). We can observe that the mutual information for the magnitude features is always higher than for the phase information features. However, the MI computed from the phase features provides evidence that information about the speaker is contained in the phase. Note that in this case, the MI is a concave function. The reason of this behavior is related to the bias and the variance of the model. In open test conditions, the performance of the recognition system (i.e., P_e and MI) improves as the number of parameters increases due to a reduction on the bias of the model. However, the variance of the model grows as well as the number of parameters increases due to the uncertainty in the parameter estimation, until the variance of the model is extremely high that has a detrimental effect on the performance of the system [16]. Moreover, we can notice an agreement between low P_e and high MI for the different feature sets. Although the best performance is achieved for MFCCs, the AMFs maintain their performance for higher number of coefficients indicating that information can be extracted from these coefficients. In figure 4, we present the performance evaluation of a SV system using the second setup. Table 1 shows the EER and the MI for the different features showing that the best performance can be achieved with the AMFs. Additionally, an agreement between the MI computed previously and the DET curves of the different features is illustrated. Note that the magnitude features provides higher MI and also better performance.

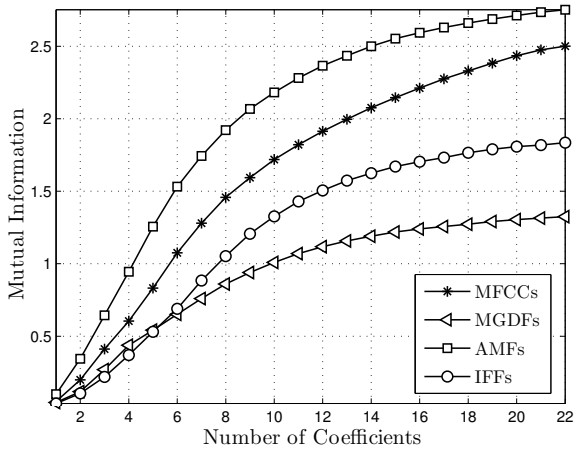

Figure 2: Comparison of MI as a function of the number of coefficients for different magnitude and phase features (*closed-test*).

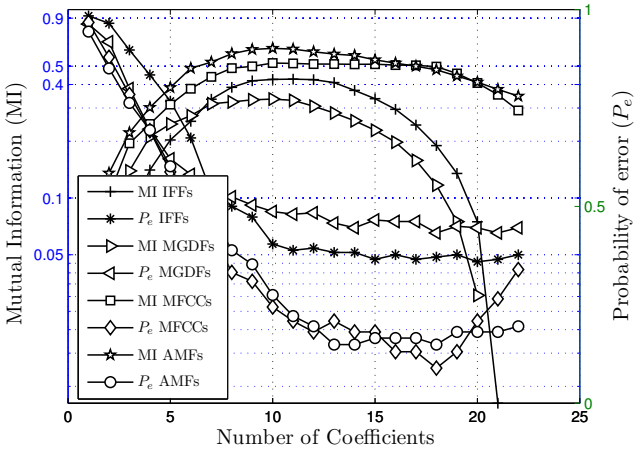

Figure 3: Comparison of MI and P_e for different magnitude and phase features (*open-test*).

8. CONCLUSION

In this work, we extracted features based on the AM-FM representation and compared with others from an informa-

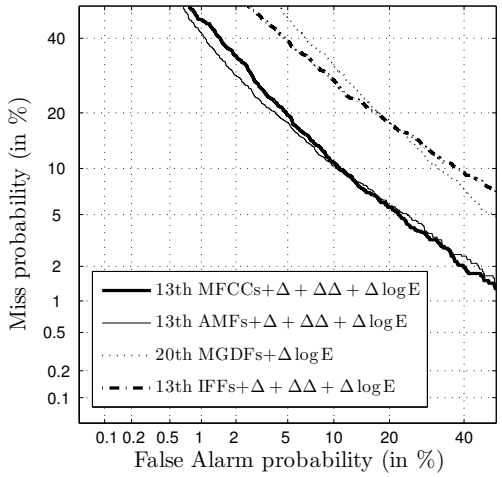

Figure 4: Comparison of DET curves for magnitude and phase features.

Table 1: EER and MI comparison for different feature sets.

Feature Set	EER %	MI
13th MFCCs +Δ+ΔΔ+Δ log E	10.4561	0.2957
13th AMFs +Δ+ΔΔ+Δ log E	10.2267	0.3064
20th MGDFs+Δ log E	19.0134	0.1262
13th IFFs +Δ+ΔΔ+Δ log E	18.7126	0.2325

tion theoretical point of view and show that the MI is closely connected to the performance of the system (P_e). Our results show that the magnitude features attain higher MI and lower P_e. We highlighted that the knowledge of the MI between features and the speakers can help us for a better feature selection. Moreover, we address the issue that the phase of speech signals contains important discriminative information, useful for speaker recognition. Finally, we present that for our SV database, the AM-FM representation attains better performance.

REFERENCES

[1] J. Campbell, "Speaker recognition: a tutorial," in *IEEE Proceedings*, vol. 85, Sept. 1997, pp. 1437–1462.

[2] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Signal Process.*, vol. 29, no. 2, pp. 254–272, 1981.

[4] M. H. Rajesh, A. M. Hema, and G. Venkata Ramana Rao, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. on Audio, Speech, and Language Process.,*, vol. 15, no. 1, pp. 190–202, 2007.

[5] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 16, no. 6, pp. 1097–1111, 2008.

[6] D. P. W. Ellis and J. A. Bilmes, "Using mutual information to design feature combinations," in *Proceedings ICSLP*, Oct. 2000, pp. 79–82.

[7] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[8] D. V. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Letters*, vol. 12, no. 9, pp. 621–624, 2005.

[9] M. Sun and R. J. Sclabassi, "Discrete-time instantaneous frequency and its computation," *IEEE Trans. on Signal Process.*, vol. 41, no. 5, pp. 1867–1880, 1993.

[10] F. B. *et al.*, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Process.*, pp. 430–451, 2004.

[11] J.-L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[12] T. Eriksson, S. Kim, H.-G. Kang, and C. Lee, "An information-theoretic perspective on feature selection in speaker recognition," *IEEE Signal Process. Letters*, vol. 12, no. 7, pp. 500–503, 2005.

[13] J. Campbell, "Testing with YOHO CD-ROM voice verification corpus," in *Proceedings ICASSP*, May 1995, pp. 341–344.

[14] "The NIST 2004 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2004/.

[15] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings Odyssey*, Jun. 2001, pp. 213–218.

[16] L. Ljung and E. J. Ljung, *System identification: theory for the user*. Prentice-Hall Englewood Cliffs, NJ, 1987.