Ambient Noise Classification for Automatic Speaker Identification

R. Santana, L. Zão and R. Coelho
Electrical Engineering Department
Instituto Militar de Engenharia (IME)
Rio de Janeiro, Brazil
email:{santana80,zao,coelho}@ime.eb.br

***Abstract*— This paper proposes two methods for acoustic ambient noises classification. The classification is based on the Kurtosis coefficient and the Bhattacharyya distance. Five colored acoustic noises, some captured in different environments and a White artificially generated, were used to perform the classification methods. These noises were obtained from NOISEX-92 database. Automatic speaker identification experiments were conducted using TIMIT speech database, corrupted with the acoustic noises. Mismatch conditions (SNR of 10 dB, 15 dB and 20 dB) were also examined in the experiments. The performances presented considerable variations among the different acoustic noises. The results show that the noise classification obtained with the proposed methods could detect the differences in the speaker identification accuracies. The MFCC (Mel-Frequency Cepstrum Coefficients) and GMM (Gaussian Mixture Models) were applied for the identification experiments.**

***Keywords*— *Automatic Speaker Recognition, ambient noises, noises classification*..**

I. Introduction

Biometric authentication [1] is based on human characteristics, such as fingerprint, iris, face and voice. The usage of such methods in access control applications is being applied in systems with security concerns [2]. The biometric solutions have many advantages in comparison to passwords and identity cards access control methods. Speech is considered the most natural biometric feature to recognize a person [3].

The speech signal conveys several levels of information such as: words, message spoken, and the identity of the speaker. Moreover, the speech features extraction is considered simple using the available technology. Automatic speaker recognition (ASkR) systems are widely used in access control, data security and forensic applications [3].

Recently, the provision of robust speaker recognition to noisy environments became an important issue [4] [5]. One of the major challenges of this area is referred to the variability of the acoustic environment noise statistics. Solutions based on missing features and multicondition training [4] were proposed to deal with this drawback. However, none of these proposals explain the different recognition performances obtained with distinct ambient noises.

This paper proposes two methods for acoustic noises classification. These methods are based on the Kurtosis coefficient (K) [6] and the Bhattacharyya distance (Bd) [7]. The Kurtosis coefficient measures ”heaviness of tails” of random processes [8]. The Bd was used as a measure of divergence among the various noisy speech signals. The speaker identification accuracy was examined considering the proposed noise classification.

The experiments were performed with five acoustic noises, obtained from the NOISEX-92 database [9]. Three different signal-to-noise ratios (SNR) were considered: 10 dB, 15 dB and 20 dB, for the mismatch conditions tests. The TIMIT speech database [10] was applied in the experiments. The results showed that the differences among the speaker identification accuracies were detected by the proposed noises classification.

The rest of this paper is organized as follows. In Section II, the speech features and classification models used in the ASkR systems are briefly described. Section III presents the Kurtosis coefficient and the Bhattacharyya distance, that were proposed for ambient noises classification. Section IV describes the TIMIT speech database and NOISEX-92 noise database. The experiments results and the noises classification are also presented in Section IV. Finally, Section V concludes the paper.

II. Automatic Speaker Recognition

A complete ASkR system can be divided into two phases: training and testing. During the training phase, or enrollment, the speaker's models are generated and stored in the system. In the testing phase, the speaker is compared to the models previously generated.

Each ASkR phase is generally composed of three consecutive steps: speech acquisition/pre-processing, features extraction and classification (see Fig. 1). In the first step, the speech signal is windowed into small time frames. For each frame, the Mel-Frequency Cepstral Coefficients (MFCC) [11] and the delta coefficients were used to compose the feature vectors. These vectors are then concatenated into a feature matrix. The classification step is responsible for generating the speaker models based on the feature matrix. During the testing phase, the feature matrix is compared to those previously generated models.

The classification step is composed of two tasks: identification and verification. In the speaker verification, the speaker makes an identity claim and the system accepts or rejects this claim. In the speaker identification, there is no identity claim, and the system decides who the person

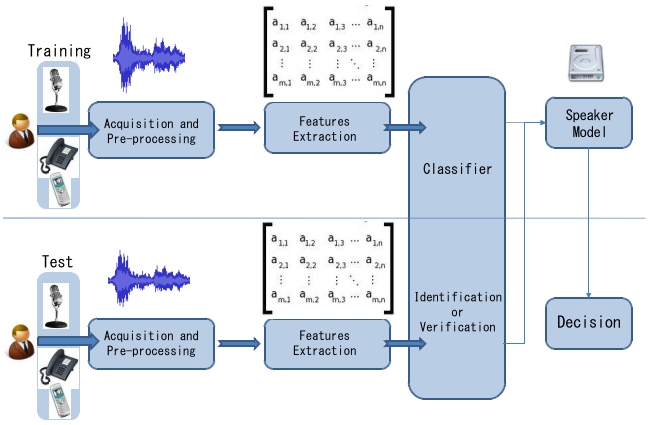

Fig. 1. The steps of an Automatic Speaker Recognition.

is from a limited set of possible speakers. The Gaussian Mixture Models (GMM) were used in this study.

The speaker recognition system can also be classified into text-dependent or text-independent. In the text-dependent speaker recognition, the system previously knows which words are being spoken by the user. In the other hand, the speaker can use any words or phrases in the text-independent case. This paper focus on the study of text-independent speaker identifications.

A. MFCC Feature

The MFCC features [12] extraction schematic is depicted in Fig. 2 [13]. These features represent the speech spectrum in a short period of time.

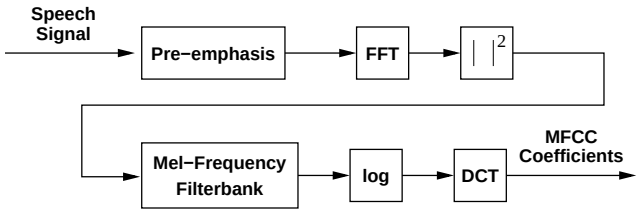

Fig. 2. Representation of the MFCC extraction.

A Mel-Frequency filterbank (Fig. 2) is understood as having linear spacing below 1000 Hz and logarithmic spacing above 1000 Hz. These frequencies can be obtained from a linear-frequency scale through the following relation:

$$F_{Mel} = 1127 \cdot \ln\left(1 + \frac{F_{Hz}}{700}\right) \tag{1}$$

The MFCC coefficients are calculated according to Eq. 2:

$$c_j = \sum_{k=1}^{M} X_k \cdot \cos\left[j\left(k - \frac{1}{2}\right)\frac{\pi}{M}\right], \quad j = 1,2,\ldots,D \tag{2}$$

where M is the number of filters in the Mel-Frequency filterbank, X_k is the log-energy output of the k^{th} filter, and D is the number of cepstrum coefficients.

For each time frame, a D-dimensional feature vector $\vec{x}$ is formed with the coefficients calculated in Eq. 2. For each speech segment, composed of T frames, the obtained feature vectors are concatenated into a $D \times T$ feature matrix. This feature matrix is then used in the classification models.

B. Delta Feature

The delta coefficients capture the dynamic information and remove the time-invariant spectral information of the feature vectors [5]. In this work, the delta coefficients are obtained as the time differences between the MFCC coefficients. Thus, for a set of MFCC feature vectors $\vec{x}_i$, the delta features are formed as follows.

$$\Delta\vec{x}_i = \vec{x}_i - \vec{x_{i-W}} \tag{3}$$

The delta coefficients are called dynamic features, while the MFCC are called static features. The dynamic features are generally used in ASkR together with the static feature vectors.

C. GMM

The GMM (λ) is composed of a weighted sum of M densities, given by

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{4}$$

where $\vec{x}$ is a random vector of dimension D, p_i, $i = 1,\ldots,M$, are the mixture weights, and $b_i(\vec{x})$, $i = 1,\ldots,M$, are the density components. Each component density is a D variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{|K_i|}}\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T K_i^{-1}(\vec{x}-\vec{\mu})\right) \tag{5}$$

with mean vector $\vec{\mu}_i$ and covariance matrix K_i, where T denotes the transpose operation and $|.|$ is the determinant.

A Gaussian Mixture Model is completely parametrized by mean vectors, covariance matrices, and mixture weights presented in Eqs. 4 and 5:

$$\lambda = \{p_i, \vec{\mu}_i, K_i\} \qquad i = 1,\ldots,M \tag{6}$$

The GMM parameters are estimated using a special case of the expectation-maximization (EM) algorithm [5]. For a feature matrix X, composed of a sequence of T independent vectors $X = \{\vec{x}_1,\ldots,\vec{x}_T\}$, the normalized log-likelihood of the GMM is given by

$$\log p(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|\lambda) \tag{7}$$

During the training phase, the model parameters are chosen as the ones that maximize the likelihood in Eq. 7. During the testing phase, the speaker identification system chooses the speaker model for which the likelihood value in Eq. 7 is maximum.

III. Noise Classification

This Section presents the proposed two different measures that were used to classify the acoustic noises: the Kurtosis coefficient [6] and the Bhattacharyya distance [7]. The results of these measures are presented in Section IV.

A. Kurtosis Coefficient

The Kurtosis coefficient of a random process measures how its values fall a long way from the mean [8]. It is defined as follows [6].

$$K = \frac{E[(X(t) - m_X)^4]}{\sigma_X^4} \tag{8}$$

where m_X and σ_X are, respectively, the mean and the standard deviation of a random process $X(t)$, and $E[.]$ means the first order moment of the dispersion.

For a Gaussian random process, we have $K = 3$. The Kurtosis is used to measure how similar a random process is from having a Gaussian distribution. If a random process has $K \approx 3$, its distribution is considered similar to a Gaussian. In the other hand, random processes with $K \neq 3$ present distributions not similar to a Gaussian one. An artificially generated White Gaussian noise is used as a reference to the Kurtosis classification results.

B. Bhattacharyya Distance

Given two random variables, X_1 and X_2, with probability density functions $p_1(x)$ and $p_2(x)$, respectively, the Bhattacharyya coefficient (ρ) is defined as

$$\rho = \int_{-\infty}^{\infty} \sqrt{p_1(x) \cdot p_2(x)} \cdot dx \tag{9}$$

From Eq. 9, it follows that

$$Bd(X_1, X_2) = -\ln\rho = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x) \cdot p_2(x)} \cdot dx \tag{10}$$

The Bd obeys the following properties:

- $0 \leq Bd < \infty$;
- $Bd(X_1, X_2) = 0 \Leftrightarrow p_1(x) = p_2(x)$ for all x.

The properties above show that, as the Bd value increases, the distance between the distributions of the random processes X_1 and X_2 also increases.

IV. Experiments

This Section presents the identification results considering various speaker identification tests, using both clean and noisy speech signals. From the five acoustic noises used for this purpose, a white Gaussian one was artificially generated. The other noise signals were captured in four different real noisy environments. The noise corruptions were done using three different SNR: 10 dB, 15 dB and 20 dB (i.e., mismatch condition experiments). The differences among the identification performances obtained for the different noises are compared and interpreted using the proposed Kurtosis and the Bhattacharyya distance classifications.

A. Speech Database

The TIMIT speech database [10] was used in the speaker identification experiments. The experiments were conducted using all 630 TIMIT speakers (438 male and 192 female). Each speaker recorded ten speech utterances with duration of about 3 seconds. Fig. 3(a) depicts the spectrogram of one of these utterances. The database was recorded at a sampling rate of 16 kHz, 1-channel PCM, and 16-bit resolution. All the 6300 utterances were recorded using the same handset.

For the identification experiments, 8 speech utterances of each speaker were concatenated to be used in the training phase. The other 2 utterances were used in the testing phase. Thus, the experiments were conducted with long-time duration for training (about 24s) and short-time duration for tests (about 3s). Each experiment had 630 speakers $\times$ 2 test utterances per speaker $=$ 1260 tests. The same configuration was used in [4].

B. Noise Database

The NOISEX-92 [9] database is originally composed of 15 different acoustic noises and it is freely available at the IEEE Signal Processing Information Base. A subset of this database was used in the experiments conducted in this study. All the noises were captured with a sampling rate of 19.98 kHz, 16-bit resolution and 235-second duration.

Tab. I describes the five ambient noises that were used to corrupt the TIMIT speech utterances. The noises were re-sampled to a 16 kHz sampling rate before being added to the speech utterances. Fig. 3(b)-(f) show the spectrogram of a speech utterance corrupted by the five acoustic noises with SNR of 10 dB.

TABLE I
Five noises extracted from the NOISEX-92 database

Noise	Description
Volvo	A car at 120 km/h, asphalt road and rainy conditions
Factory	Noise recorded in a car production hall
M109	A M109 tank moving at a speed of 30 km/h
White	Artificially generated white Gaussian noise
Destroyer	Noise recorded in the engine room of a Destroyer

C. Speaker Identification Accuracy Results

For the identification experiments, each speech utterance was divided into frames with duration of 20 ms. The frames were obtained using Hamming windows with 50% superposition. The feature matrices were primarily composed of 12-dimensional MFCC vectors, obtained from 26 filters (see Eq. 2).

The GMM speaker models were obtained using 32 Gaussians. The identification tests were also performed with clean speech signals. Tab. II shows the results obtained from these identification scenarios.

It can be seen from Tab. II that the identification accuracies were very affected by the acoustic noises. While clean speech tests presented 93.89% accuracy, the identification results in noisy environment achieved from 20.53% to 42.62% in average. Moreover, the noise sources led to very

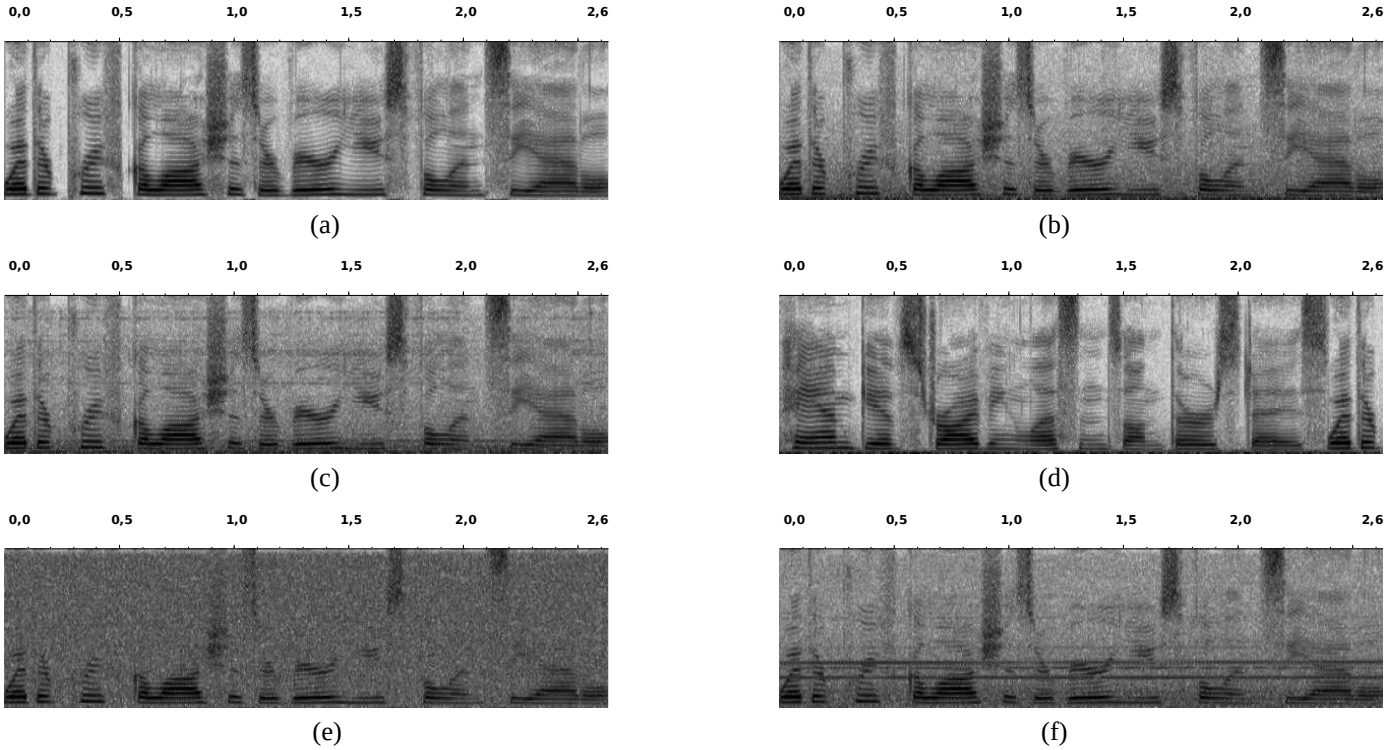

Fig. 3. Spectrogram of a TIMIT speech utterance corrupted by the each of the acoustic noises: (a) clean; (b) *Volvo*; (c) *Factory*; (d) *M109*; (e) *White* and (f) *Destroyer*.

TABLE II
IDENTIFICATION ACCURACY(%) FOR 12-DIMENSIONAL MFCC FEATURE VECTORS

Noise	SNR			Average
	10 dB	15 dB	20 dB	
Clean	93.89			
Volvo	26.19	40.79	60.87	42.62
Factory	17.06	38.10	60.63	38.60
M109	10.87	30.87	54.76	32.17
White	10.56	28.49	57.14	32.06
Destroyer	5.16	15.95	40.48	20.53
Average	13.97	30.84	54.78	33.20

TABLE III
IDENTIFICATION ACCURACY(%) FOR 24-DIMENSIONAL MFCC + DELTA FEATURE VECTORS

Noise	SNR			Average
	10 dB	15 dB	20 dB	
Volvo	31.03	47.86	66.98	48.62
Factory	19.52	40.56	62.78	40.95
M109	12.14	32.14	57.22	33.83
White	10.79	27.86	55.00	31.22
Destroyer	5.56	17.62	45.32	22.83
Average	15.81	33.21	57.46	35.49

different identification results. The worst result was achieved with the *Destroyer* noise, with SNR of 10 dB. This means a reduction of 88.73% in the identification accuracy. The results shown in Tab. II are presented according to the noisy experiments performances (from the best accuracies on the top to the worst results on the bottom). The greatest differences between noisy performances were achieved with *Volvo* and *Destroyer* noises. This difference was about 25% with SNR of 15 dB.

In order to achieve better recognition results under noisy conditions, the experiments were repeated including the delta features vectors (12 MFCC + 12 delta coefficients). Tab. III presents the results obtained for these experiments.

Note that, in general, the delta coefficients improved the performances in comparison to the use of single MFCC coefficients. The average accuracy increased from 33.20%, with 12 MFCC, to 35.49%, with 12 MFCC + 12 Delta. While the *Volvo* noise presented an average improvement of 6.00%, the obtained with *M109* was only about 1.66%. However, for the *White* noise, the delta coefficients did not improve the identification accuracies. In this case, the average results decreased from 32.06% to 31.22%. It can also be seen that the greatest accuracy difference in the results of the noisy tests achieved more than 30% (47.86% with *Volvo* and 17.62% with *Destroyer*) with SNR of 15 dB.

D. Noise Classification

This Section presents the noise classification results. The proposed noises classification is used to explain the differences in the speaker identification performances.

1) Kurtosis: Tab. IV presents the Kurtosis coefficients (K) for the five acoustic noises.

Considering the obtained K values, the noises were classified into three categories: noises with $K \approx 3$ (Gaussian distribution); $K < 3$ and $K > 3$. As expected, the Kurtosis

TABLE IV
THE KURTOSIS MEASURES FOR THE FIVE ACOUSTIC NOISES

Noise	Kurtosis Ratio (K)
Volvo	3.445
Factory	3.097
M109	2.959
White	2.984
Destroyer	2.870

value of *White* noise was $K \approx 3$, since it was artificially generated with a Gaussian distribution. The identification results with *Factory* and *M109* were quite similar to those with *White*. This is due to the fact that they presented $K \approx 3$. *Volvo* noise, with $K > 3$, presented the best identification accuracies. *Destroyer* noise presented $K < 3$ and the worst results.

2) Bhattacharyya distance: Five randomly selected speech utterances were used for the *Bd* evaluation. These utterances were spoken by five different speakers (3 male and 2 female). The *Bd* was evaluated for each pair of clean speech utterances. The measures were also obtained using the speech utterances corrupted by the different noises with SNR of 10 dB, 15 dB and 20 dB, keeping the same five speakers. Tab. V presents the mean values of the *Bd* for the noisy and clean utterances.

TABLE V
THE MEAN VALUES OF THE BHATTACHARYYA DISTANCES FOR FIVE SPEECH UTTERANCES AND DIFFERENT NOISE CORRUPTIONS (SNR OF 10 DB, 15 DB AND 20 DB)

Noise	SNR		
	10 dB	15 dB	20 dB
Clean	0.0326		
Volvo	0.0298	0.0297	0.0304
Factory	0.0278	0.0283	0.0292
M109	0.0263	0.0268	0.0279
White	0.0256	0.0268	0.0280
Destroyer	0.0261	0.0270	0.0284

The results show that the *Bd* measured for the noisy speech signals, were lower than the *Bd* value of the clean speech signals. This means that the speech utterances become closer (smaller distances) when corrupted by the ambient noises. It can be noted that *Volvo* noise, that presented the greatest *Bd* values, has also the best identification accuracies. This relationship was also achieved for the *Factory*, *M109* and *White* noises. This means that the higher the *Bd* values, the best accuracies. *Bd* values for *Destroyer* noise were not interesting to explain the identification accuracy.

Although the *Bd* did not present the relationship between its values and the identification results for all the acoustic noises, it could be used together with the Kurtosis coefficient for the noises classification. So, the classification based on the Kurtosis coefficient and the Bhattacharyya distance were able to detect the variable speaker identification accuracies due to the ambient noises.

V. CONCLUSION

Two methods for ambient noises classification were proposed in this paper. They are based on the Kurtosis coefficient and the Bhattacharyya distance. The methods were applied to five different acoustic noises. Identification experiments were conducted using the MFCC (12-dimensional) and MFCC+Delta (24-dimensional) feature vectors, and the GMM classifier. In the experiments, the acoustic noises were added to the speech database. Three different mismatch conditions with SNR 10 dB, 15 dB and 20 dB were also examined during the tests.

The speaker identification results showed that the acoustic noises led to severe performance degradation. Moreover, different noises presented a great variability in the identification results. Some of these differences in the experiments performances achieved more than 30%.

The classification with Kurtosis enabled the definition of three different noise classes. Noises within the same class presented similar identification results. The Bhattacharyya distance results showed that the best identification values were achieved for the highest *Bd* values. The proposed classification showed to be very promising to classify acoustic noises. Moreover, it enables to understand the noises impact on the speaker identification accuracies.

REFERENCES

[1] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[2] R. Kuhn, M. Tracy, and S. Frankel, "Security for telecommuting and broadband communications," *NIST Recommendations*, vol. 800-46, pp. 1–113, August 2002.

[3] J. Campbell, J.P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437 –1462, sep 1997.

[4] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 1711–1723, July 2007.

[5] D. Reynolds and R. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 72–82, 1995.

[6] M. G. Bulmer, *Principle of Statistics*. New York: Dover Publications, 1967.

[7] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *Communication Technology, IEEE Transactions on*, vol. 15, pp. 52 –60, february 1967.

[8] A. O. Allen, *Probability, Statistics, and Queueing Theory with Computer Science Applications*. Orlando, FL, USA: Academic Press, Inc., 1978.

[9] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition ii: Noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.

[11] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, vol. 8, pp. 93 – 96, apr 1983.

[12] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, pp. 357 – 366, aug 1980.

[13] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, M. I. Chagnolleau, S. Meignier, T. Merlin, O. J. Garcia, P. Delacretaz, and Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.