Automatic Speaker Recognition Based on Mel-Frequency Cepstral Coefficients and Gaussian Mixture Models

SHEERAZ MEMON*, SANIA BHATTI**, AND FARZANA RAUF ABRO***

RECEIVED ON 18.12.2011 ACCEPTED ON 21.06.2012

ABSTRACT

This paper investigates the task of SR (Speaker Recognition) for the state-of-the-art techniques. The paper initially presents the technical description of automatic SR, followed by the comparative analysis of a number of methods available for feature extraction and modeling. Based on this analysis the NIST 2001, NIST 2002, NIST 2004 and NIST 2006 Speaker recognition corpora are used to investigate the state of the art feature extraction and modeling techniques. The state of the art technique for feature extraction is delta MFCC (Mel Frequency Cepstral Coefficients) and for modeling is GMM (Gaussian Mixture Models) based on EM (Expectation Maximization). Further in this paper the details about the enrollment/training and recognition/testing is also presented. For different stages of SR systems the conventional methods are summarized.

Key Words: **Mel Frequency Cepstral Coefficients, Gaussian Mixture Models, Expectation Maximization, Speaker Recognition.**

1. INTRODUCTION

The task of identifying an individual from voice is called SR. Since early 1970's the task of speaker recognition is under investigation. It is often divided into two related applications, speaker identification and speaker verification. Establishing the identity from a list of potential candidates is called identification and accepting or rejecting a claim of an individual identity is called speaker verification. Speaker recognition may be further categorized into text-independent and text-dependent recognition. In text-dependent system the text of utterances must be same for enrolment and recognition. In text-dependent recognition, the text is taken unique for all speakers such as a unique pass phrase. Text-independent systems are most often used for speaker identification. In this case the text during enrolment and identification can be different.

In this paper the experiments based on delta-MFCC and GMM-EM are performed for the task of text-independent speaker verification and text-dependent speaker identification.

SR system shown in Fig.1 works in two stages enrolment/ training and recognition/testing. The features are extracted from speech utterances at enrolment stage [1,2]. The stochastic models are generated from characteristic feature vectors. Since the generation of stochastic models takes a

* Associate Professor, Department of Computer Systems Engineering, Mehran University of Engineering & Technology, Jamshoro.
** Associate Professor, Department of Software Engineering, Mehran University of Engineering & Technology, Jamshoro.
*** Assistant Professor, Department of Electronic Engineering, Mehran University of Engineering & Technology, Jamshoro.

reasonable amount of speech utterances in consideration, thus requires a substantial amount of time, due to this reason, training is always performed offline, and repeated only if the models do not remain valid. The training process is elaborated in detail in Fig. 2.

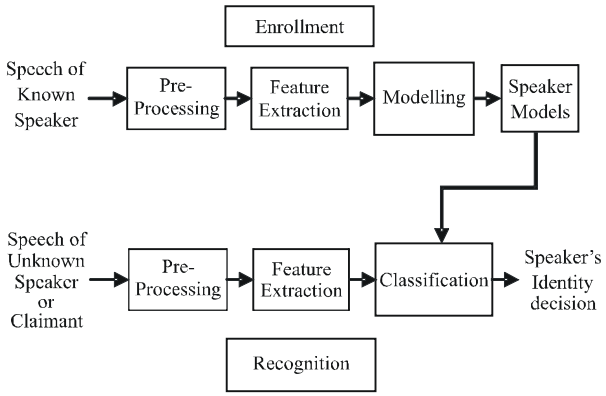

FIG. 1. OVERVIEW OF A SPEAKER RECOGNITION SYSTEM

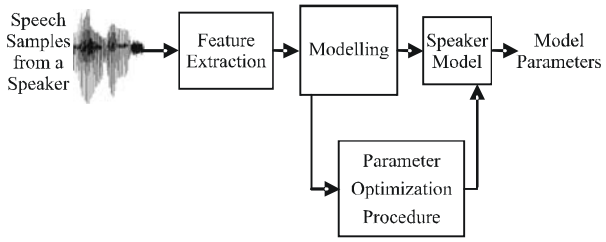

FIG. 2. ENROLLMENT OR TRAINING PHASE

In recognition stage is performed after the training stochastic models are created. In this stage the speech utterance different then used in training are utilized [1,2]. Similar feature extractor as used in training is applied to calculate the speaker specific features, thus a classifier then matches the test template to the training phase stochastic models and calculates the similarity index. This process ends with the decision in which it is determined that weather to accept or reject the claimant identity [3-6]. The recognition stage takes less amount of time and mostly performed online. The recognition stage of typical speaker identification task is given in Figs. 3-4 shows the recognition stage of a typical speaker verification system.

2. PRE-PROCESSING AND SPEECH ANALYSIS

The removal of noise, channel distortion, lip-radiation, silence and unvoiced speech is achieved in pre-processing stage [4-7]. These stages add alot towards improving the SR results. The short time signal analysis

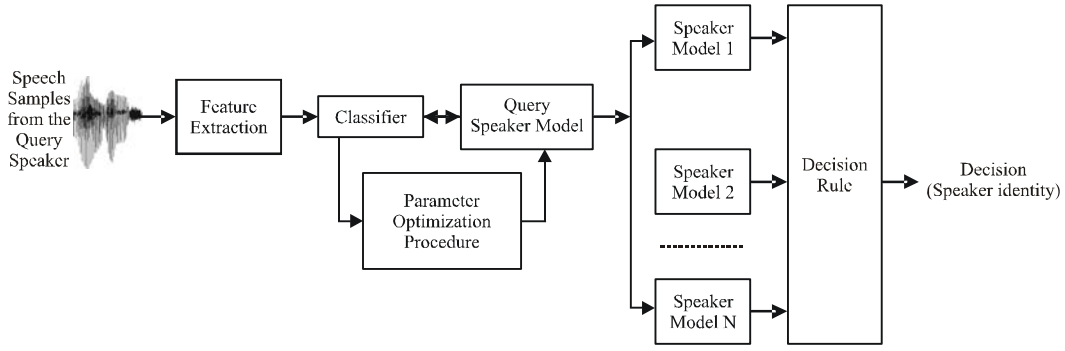

FIG. 3 TESTING PHASE FOR SPEAKER IDENTIFICATION

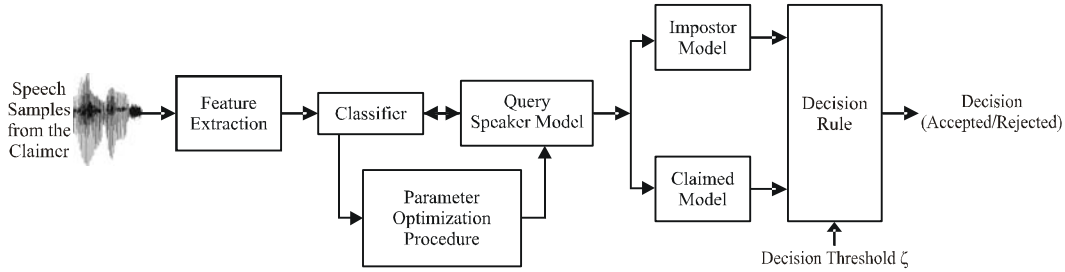

FIG. 4 TESTING PHASE FOR SPEAKER VERIFICATION

is also performed in this stage, where the speech frames are divided into size of 2-5ms and the windowing is performed. These short frames are then used to perform feature extraction. During short time signal analysis the length of the analysis window depends upon the feature extractor being used [8].

The vocal tract information available in the speech can be best determined by analyzing the speech frames of size 10-30ms [9-12]. To demonstrate the experiments in this paper we have also utilized frames of size 10-30ms. The feature information available in excitation source can be best determined by using speech frames of size 3-5ms [13]. In order to extract the information available in behavioral traits the speech frames of size 100-300ms is used.

3. FEATURE EXTRACTION

The process of converting a raw speech signal into a sequence of acoustic feature vectors carrying characteristics information about the speaker is called feature extraction. Majority of current feature extraction methods in speaker recognition use parameters derived from the classical source-filter theory. The classical source-filter theory about voice activity assumes that when air flows through vocal folds (source) and vocal-tract (filter), its flow is unidirectional. During phonation, the vocal folds vibrate. A single vibration cycle comprises of opening phase and closing phase, moving vocal folds apart or together, respectively. The frequency of vibration is determined by number of cycles per second. This frequency is subjectively perceived as pitch or objectively measured as fundamental frequency F0. The sound is then modulated by the vocal tract configuration and the resonant frequencies of the vocal tract, known as formants. Finally the speech signal is passed through the low-pass lip radiation filter which reduces the signal energies with frequency by about 6dB/octave [14]. Thus the uniqueness of the speaker specific information may be because of a number of factors such as vocal tract's shape and size, dynamics of the articulators, rate of vibration of vocal folds, the accent of the speaker and finally the speaking rate. All these factors are reflected in the speech signal, and hence are useful for SR. A number of studies have been takes in order to analyze the variability in speech signal [15], however no general conclusion have been marked to state "what constitutes a voice print". However, these studies have resulted in a variety of methods used to perform feature extraction. Thus speech features are classified based on the domain in which the analysis is conducted [16,17], the characteristic features can be divided into:

- Spectral Features: Based on short-time spectrum.
- Dynamic Features: Time derivatives of spectral features.
- Prosodic Features: Features extracted using fundamental frequency etc.

The prosodic features are further classified based on frame length into following types:

- Source Features: These features are limited to single glottal period.
- Suprasegmental Features: These features span over few glottal periods.
- High-Level Features: These features span over a word or utterance.

For this paper we have used the spectral features called MFCCs, with its dynamic attribute called delta features.

The MFCCs best describe a speaker model [9,18] because psychophysical studies have established the fact that

perceiving of frequency content from speech pursue a scale, which is non linear in nature called Mel scale [17,19], given as:

$$f_{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

Where fmel denotes the subjective pitch in Mels, twelve MFCCs are derived to generate a speaker model.

The features which represent time derivatives of the spectrum-based features are referred to as the dynamic features. Dynamic cepstral features such as delta (first derivative of cepstral features) and double-delta (second derivative of cepstral features) have played an important role to capture transitional characteristics of voice [12]. Twelve delta and twelve double delta coefficients are then obtained from twelve MFCCs, thus a 36 dimensional feature vector is used to model speakers from NIST speech corpora.

4. SPEAKER MODELLING

The modeling techniques generate the stochastic models of speech features. The objective of modeling methods is to produce a speaker model, with unique representation, using feature vectors as inputs.

The template matching techniques [20] were the most widely used techniques for SR at the early stages of this technology. Template matching uses training and testing vectors, evaluated using similarity measure. The most common techniques used in this regard are, Mahalanobis or Euclidean distance, and spectral distance.

GMM [7,21], HMM (Hidden Markov Models) [22], SVM (Support Vector Machines) [23], VQ (Vector Quantization) [21,24], and ANN (Artificial Neural Networks) [25] are the modern classifiers used in speaker recognition technology. GMM has recently been turned as most widely used modeling technique for SR [7]. Thus we also use GMM based on EM to perform speaker modeling.

4.1 GMM-EM

The GMM [23] is a feature modelling and classification algorithm widely used in the speech-based pattern recognition, since it can smoothly approximate a wide variety of density distributions. The adapted GMM [2] which consists of UBM (Universal Background Model) based on MAP (Maximum A Posteriori) estimation have turned GMMs into reality. GMMs use EM algorithm for the optimization of GMM parameters such as means, covariances and weights.

The GMM models the PDF (Probability Density Function) of a feature set given as:

$$p(x \mid \lambda) = \sum_{t=1}^{M} p_i b_i(x) \quad (2)$$

Here x refers to a D-dimensional random vector. For every Gaussian mixture component the defining formula is:

$$b_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \quad (3)$$

In Equation (3) μi and Σ denote the mean and covariance respectively. The weights assigned to each component complies the condition that sum of all weights should be equal to one, denoted as, $\sum_{i=1}^{M} p_i = 1$. Thus a speaker model is denoted as:

$$\lambda = \{p_i, \mu_i, \Sigma_i\}, i=1,.....,M \quad (4)$$

5. EXPERIMENTS

5.1 Speech Corpora

The speech data comes from various commercial and academic sources and has been produced for a wide variety

of applications and developed under different conditions. Although, in the recent years the NIST database [26,27] has become the most frequently used corpora, other data sets are still being used as they can provide performance evaluation across different recording environments, populations of speakers and different languages. NIST launches an evaluation plan every year and as a result provides a speech corpus to perform the experiments. We have used NIST 2001, NIST 2002, NIST 2004, and NIST 2006 evaluation speech corpora, to perform the experiments.

5.2 Overview of Speaker Verification System

The SR system used to demonstrate the experiments is shown in Fig. 5. The recognition is performed after configuring the system in three possible stages. The first stage is MAP based UBM training stage, in this stage NIST 2001 and NIST 2002 speech corpora are used to obtain the UBM, approximately one hour of speech is used to obtain this model, about 1024 GMM components are formed. In second stage the training is performed, in order to perform the text-independent speaker verification, NIST 2004 speech corpus is used. For text dependent speaker identification task the NIST 2006 speech corpus is utilized. About 5 min of training speech is used to generate speaker models and the target speaker means are then adjusted away from the UBM using MAP

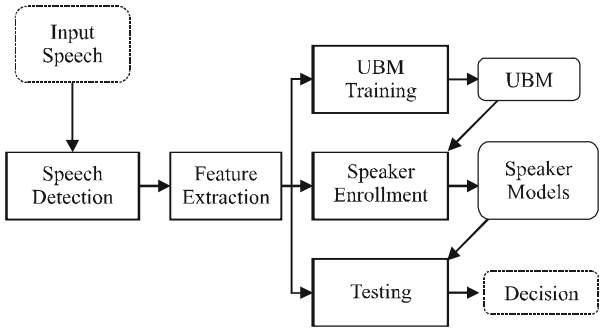

FIG. 5. MAJOR STAGES OF EXPERIMENTAL SETUP

estimation. The third stage is of testing or validation; about 5min length test utterances from the respective corpora are used to evaluate the performance of text-independent speaker verification and text dependent speaker identification.

For the above stages the speech detection and feature extraction methods used are identical. For speech detection an energy based silence detector which identifies the low energy portions of the signal [27] as silence regions is used. Previous research [9] have shown that MFCC based system is relatively robust to the changes in frame size (in the range 20-50ms) and frame step (in the range 1/6 to 1/3 of the frame size). Thus, we employed MFCC to characterize the speaker information. The feature vector representing a given frame had 36 dimensions including: 12 MFCCs, 12 velocity coefficients (ΔMFCC-first derivative of MFCC), and 12 acceleration coefficients (ΔΔMFCC-second derivative of MFCC). MAP-UBM based GMM is then used to model the sequence of feature vectors.

5.3 Performance Evaluation

In order to evaluate the performance of SR system, some matching and error calculation mechanism is needed. To apply the speaker recognition system to real time applications the validation of the algorithms needs to be done on a repetitive validation of the same speaker. Thus calculating a number of time the impostor and true speaker scores needs a mechanism which can relate these errors. DET (Detection Error Tradeoff) curves are widely used for this purpose [28]. Before DET plots ROC (Receiver Operating Characteristic) curves [2] were in use.

DET plots two types of evaluation errors called FA (False Alarm) probability and the FR (False Rejection) probability [2,29]. As the name indicates a false acceptance error takes place when system accepts a claimant, which is an impostor. A false rejection takes place when the system rejects a true speaker as impostor. EER (Equal Error Rate) is used to

state the SR performance; it is the value at which FA probability becomes equal to FR probability. The smaller is the EER for a given speaker verification system, the better is the system performance.

As stated above the SR can either be verification or identification, therefore both types of speaker recognition are tested in order to validate the methods. In Fig. 6. Text-independent speaker verification results are summarized. The UBM is trained using NIST2001 and NIST 2002 speech corpora by using about one hour of speech. The speaker verification experiments are then performed using NIST 2004 SRE (Speaker Recognition Evaluation) corpus following the train/test rules specified in NIST 2004 corpus. The experiments are performed for varying number of Gaussian components, as 128, 512, 1024, 2048. Of which the best results are achieved for 1024 Gaussian components. In Fig.7. the results for the task of text-dependent speaker identification are summarized. The speaker identification experiments are performed using NIST 2006 SRE corpus. The description of the train/test experiments as outlined in NIST 2006 is applied. The best identification scores were obtained for 1024 Gaussian components as outlined in Fig. 7.

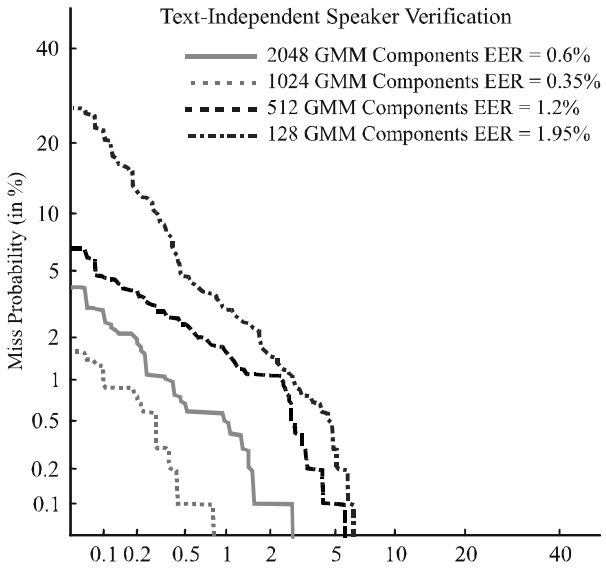

FIG. 6. EERS FOR SPEAKER VERIFICATION

6. CONCLUSION

In this paper state-of-the-art methods for the task of speaker recognition are tested. The results indicate that a UBM-GMM comprising of 1024 components yields better results, this is because the number of components above this cause thinner distribution of training data and the number of components below this yield thinner number of vectors to model a speaker. This paper also demonstrates the framework for SR system and its major stages. The methods used at pre-processing, feature extraction, modeling and performance evaluation of the SR process are explained.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Dr. Abul Qadeer Khan Rajput, Vice-Chancellor, Prof. Dr. Muhammad Aslam Uqaili, Pro Vice-Chancellor, and Prof. Dr. Mukhtair Ali Unar, Director, Institute of Information & Communication Technologies, Mehran University of Engineering & Technology, Jamshoro, Pakistan, for their efforts towards strengtheny research and development trends. This research would not have been possible without the PC-I Scheme for Ph.D. scholarship.

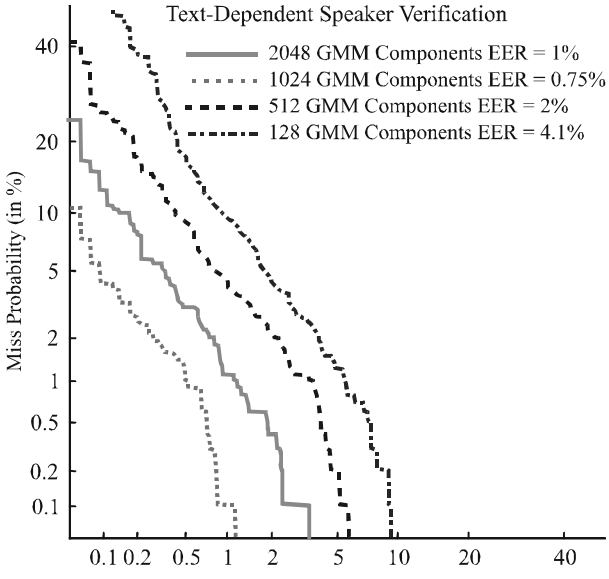

FIG. 7. EERS FOR SPEAKER IDENTIFICATION

REFERENCES

[1] Naik, J.M., "Speaker Verification: A Tutorial", IEEE Communications Magazine, 1990.

[2] Campbell, J.P., "Speaker Recognition: A Tutorial", IEEE Proceedings, Volume 85, No. 9, pp. 1437-1462, September, 1997.

[3] Furui, S., "Recent Advances in Speaker Recognition", Pattern Recognition Letters, Volume 18, No. 9, pp. 859-872, 1997.

[4] Reynolds, D.A., "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", Speech Communication, Volume 17, pp. 91-108, 1995.

[5] Sivakumaran, P., Fortuna, J., and Ariyaeeinia, A., "Score Normalization Applied to Open-Set, Text-Independent Speaker Identification" Proceedings of 8th European Conference on Speech Communication and Technology, pp. 2669-2672, Eurospeech, 2003.

[6] Kinnunen, T.H., "Optimizing Spectral Feature Based Text-Independent Speaker Recognition", Ph.D. Dissertation, University of Joensuu, Finland, 2005.

[7] Reynolds, D.A., and Rose, R., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, Volume 3, 1995.

[8] Jayanna, H.S., and Mahadeva, P.S.R., "Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition", IETE Technical Review, Volume 26, No. 3, pp. 181-190, 2009.

[9] Reynolds, D.A., "Experimental Evaluation of Features for Robust Speaker Identification", IEEE Transactions on Speech Audio Process, Volume 2, No. 4, pp. 639-43, October, 1994.

[10] Atal, B.S., "Automatic Speaker Recognition Based on Pitch Contours", Journal of Acoustical Society of America, Volume 52, No. 6, Part-2, pp. 1687-97, 1972.

[11] Rabiner, L., and Juang, B.H., "Fundamentals of Speech Recognition", Pearson Education, Singapore, 1993.

[12] Liu, Y., Russell, M., and Carey, M., "The Role of Dynamic Features in Text-Dependent and Independent Speaker Verification", IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 1, May, 2006.

[13] Satyanarayana, P., "Short Segment Analysis of Speech for Enhancement", Ph.D. Dissertation, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, February, 1999.

[14] Quatieri, T.F., "Discrete-Time Speech Signal Processing Principles and Practice", Prentice-Hall Signal Processing Series, 2002.

[15] Plumpe, M.D., Quatieri, T.F., and Reynolds, D.A., "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification", IEEE Transactions on Speech Audio Process, Volume 7, No. 5, pp. 569-85, 1999.

[16] Jain, A.K., Ross, A., and Prabhakar, S., "An Introduction to Biometric Recognition", IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image and Video Based Biometrics, Volume 14, No. 1, January, 2004.

[17] Ganchev, T.D., "Speaker Recognition", Ph.D. Dissertation, University of Patras, Greece, 2005.

[18] Memon, S., and Lech, M., "Speaker Verification Based on Information Theoretic Vector Quantization", Proceedings of IMTIC, Communication in Computer and Information Science Series, Volume 20, pp. 391-399, Springer Berlin Heidelberg, April, 2008.

[19] Memon, S., and Lech, M., "Using Information Theoretic Vector Quantization for GMM Based Speaker Verification", Proceedings of EUSIPCO, Eurasip Library, Lausanne Switzerland, August, 2008.

[20] Atal, B.S., "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", Journal of the Acoustical Society of America, Volume 55, No. 6, pp. 1304-1312, 1974.

[21] Memon, S., Khanzada, T.J.S., and Bhatti, S., "Text-Independent Speaker Verification Based on Information Theoretic Learning", Mehran University Research Journal of Engineering & Technology, Volume 30, No. 3, Jamshoro, Pakistan, July, 2011.

[22] Yuk, C.C.Q.L.D.S., "An HMM Approach to Text Independent Speaker Verification", IEEE International Conference on Acoustics, Speech and Signal Processing, 1996.

[23] Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E., and Torres-Carrasquillo, P.A., "Support Vector Machines for Speaker and Language Recognition", Computer Speech and Language, Volume 20, pp. 210-29, 2006.

[24] Soong, F.K., et al., "A Vector Quantization Approach to Speaker Recognition", AT&T Technical Journal, Volume 66, No. 2, pp. 14-26, 1987.

[25] Wang, C., Xu, D., and Jose, C.P., "Speaker Verification and Identification Using Gamma Neural Networks", International Conference on Neural Networks, 1997.

[26] NIST, “Speaker Recognition Evaluation”, 2001 http://www.itl.nist.gov/iad/mig/tests/spk/2001/

[27] Reynolds, D.A., Rose, R.C., and Smith, M.J.T., "PC-Based TMS320C30 Implementation of the Gaussian Mixture Model Text-Independent Speaker Recognition System", Proceedings of the International Conference on Signal Processing Applications and Technology, pp. 967-973, November, 1992.

[28] Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybock, M., "The DET Curve in Assessment of Detection Task Performance", EUROSPEECH, pp. 1895-1898, 1997.