SPEAKER VERIFICATION SYSTEM THROUGH TELEPHONE CHANNEL
An Integrated System for Telephony Plataform Asterisk

Alexandre Maciel, Weber Campos, Clêunio França and Edson Carvalho
Informatics Center, Federal University of Pernambuco, Recife, Brazil

Keywords: Speaker Verification, Gaussian Mixture Models, Asterisk.

Abstract: Presented in this work a system of speaker verification based in GMM approach to integration with Asterisk telephony platforms. Has developed a database itself with 30 speakers and tested for acceptance and rejection. The results were effective with success rates of 90%.

1 INTRODUCTION

The interest in using speech to verify speaker identity has increased considerably. This occurs because two factors: First, speech, opposed to others biometrics such as fingerprints and face recognition, allows recognition to be performed remotely as it can easily transmitted over communication channels. Second factor is linked to operation's quantity, value and complexity increase performed by call center systems in last years.

The speaker verification systems development through telephone channel has been widely studied in academy for a long time. Questions like microphone variability, transmission noise and environment are studied in (Naik, 1989) and (Mak, 2002) and several solutions presented with success. However, the speaker verification systems integration with telephony platforms is little studied yet. One reason is the proprietary technologies used in most of systems.

In this paper we describe a speaker verification system based on GMM (Gaussian Mixture Models) approach for telephony environment integrated with Asterisk. A speaker database was created and false acceptance rate and false rejection rate were measured for 8, 16 and 32 gaussian mixtures. O paper is organized as follows: Section 2 shows Asterisk description and architecture overview, section 3 shows speaker verification fundamentals, section 4 shows the experiments performed. Finally section 5 brings the authors conclusions and considerations.

2 ASTERISK

Asterisk is the world's leading open source telephony engine and toolkit. Offering flexibility unheard of in the world of proprietary communications, Asterisk empowers developers and integrators to create advanced communication solutions for free.

Asterisk offers resources like: voice mail, conference, IVR, and automatic call distribution. Originally designed for Linux, Asterisk also runs on different operating systems including NetBSD, OpenBSD, FreeBSD, Mac OS X, and Solaris. A port to Microsoft Windows is known as AsteriskWin32.

Asterisk's architecture is very simple. Essentially Asterisk acts as middleware, connecting telephony technologies on the bottom to telephony applications on top, creating a consistent environment for deploying a mixed telephony environment. More information about the architecture can be found at (Meggelen, 2007).

Asterisk's core contains several engines that each plays a critical role in the software's operation. The Asterisk Gateway Interface, or AGI, provides a standard interface by which external programs may control the Asterisk dialplan. Usually, AGI scripts are used to do advanced logic, communicate with relational databases and access other external resources. For Java applications, the easiest way to interact with Asterisk is via the FastAGI protocol. FastAGI basically provides a container that receives connections from the Asterisk server, parses the request and calls for scripts mapped to the called

Maciel A., Campos W., França C. and Carvalho E. (2009).
SPEAKER VERIFICATION SYSTEM THROUGH TELEPHONE CHANNEL - An Integrated System for Telephony Plataform Asterisk.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 90-93
DOI: 10.5220/0002237900900093

URL (Spencer, 2003).

3 SPEAKER RECOGNITION

Speaker recognition is a biometric modality that uses individual's voice for recognize their identity. It is a different technology from speech recognition, which recognizes words as they are articulated, which is not biometrics.

The speech signal is produced as a result of a lot of changes that occur at different levels: semantic, linguistic, articulatory and acoustic. The differences in these transformations appear as differences in acoustic properties of speech signal. Differences related to the speakers are the result of a combination of anatomical differences inherent in the vocal tract (inherent characteristics) and those related to the dynamic movement of the vocal tract, ie how the person speaking (educated characteristics). In speaker recognition, all these differences can be used to discriminate between the speakers themselves (Campbell, 1997).

A speaker recognition system consists of mainly two parts as shown in Figure 1. In the front-end is the feature generation part and in the back-end is the classification engine. During the enrollment phase the switch is turned towards the upper route and the system generates models, and during testing or evaluation phase the models are used to check a speaker identity from an input speech signal (Mashao, 2006).

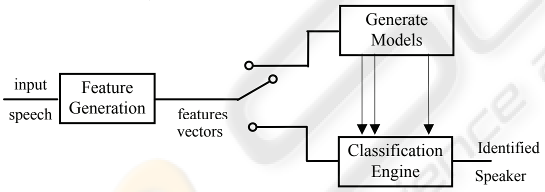

Figure 1: Speaker Recognition System.

3.1 Front-End Processing

Several processing steps occur in the front-end analysis. First, the speech signal is acquired in the time domain via sampling and after the application of the discrete Fourier transform (DFT) it is converted into the frequency domain. Then is applied a log magnitude (taking the logarithms of the magnitude of a complex signal) of the frequency domain signal (the spectrum) is computed. The signal is now said to be in the cepstral (a play on the words spectral) domain.

The signal in the cepstral domain is measured in *quefrencies* (again a play on the words frequency). The low-order *quefrencies* contain information that is due to the speech formants and therefore carries information about what is being said and the high *quefrencies* are due to the pitch and therefore assumed to be speaker dependent.

To obtain the MFC coefficients, the speech signal is windowed and converted into the frequency domain by using the DFT. In the frequency domain a log magnitude of the complex signal is obtained. A mel-scaling or mel-warping is then performed using filters. The common method of implementing these filters is to use triangular filters that are linear spaced from 0 to 1 kHz and then non-linearly placed according to the mel-scaling approximations.

The MFCC coefficient has been used for several years since the late 1990s and has successfully replaced the linear prediction cepstral coefficients (LPCC). The main advantage of the LPCC was computation but with increasing computing power and better performance (and the use of the fast Fourier transform-FFT), the MFCC has completely dominated the designs of the front-ends of speech technology systems (Mashao, 2006).

3.2 Classification Engine

For many years researches on classification engine for speech recognition has been done and some of them have reached high performance level. Many techniques have been proposed including dynamic time wrapping (DTW), Hidden Markov models (HMM), artificial neural networks (ANN) vector quantization (VQ) and Gaussian Mixture Models (GMM) (Fechine, 2000).

GMM has been used to characterize speaker's voice in the form of probabilistic model. It has been reported that the GMM approach outperforms other classical methods for text-independent speaker recognition (Reynolds, 1995) and other works has shown that the GMM also performs well for text-dependent speaker recognition (Reynolds, 2000).

For a feature vector denoted as x^t belonging to a specific speaker *s*, the GMM is a linear combination of *K* Gaussian components as follows:

$$P(x_t|\lambda_s) = \sum_{k=1}^{K} \omega_{s,k} P(\mathbf{x}_t|\mathbf{m}_{s,k},\mathbf{\Sigma}_{s,k}), \qquad (1)$$

where $\mathrm{P}(x_t|\mathbf{m}_{\mathrm{s,k}},\Sigma_{\mathrm{s,k}})$ is a Gaussian component parameterized by a mean vector $\mathrm{m}_{\mathrm{s,k}}$ and covariance matrix $\Sigma_{\mathrm{s,k}}.\omega_{\mathrm{s,k}}$ is a linear combination coefficient for speaker s (s=1, 2,…,S). Usually, a diagonal covariance matrix is used in Eq. (1). Given a sequence of feature vectors, {x1, x2, …, x_{t}}, from a specific speaker's utterances, parameter estimation for $\lambda_s = (\omega_{s,k}, \mathbf{m}_{s,k}, \Sigma_{s,k})$ (k =1,…, K; s =1,…,S) is performed by the Expectation–Maximization (EM) algorithm. Thus, a specific speaker model is built through finding proper parameters in the GMM based on the speaker's own feature vectors.

4 EXPERIMENTS

4.1 Speakers Database

The experiments described in this section were conducted with the objective of integrating the Asterisk telephony platform to an application for verification of speaker using the GMM approach. With this propose, it was developed a database with a small amount of samples for training and testing. Chosen by a small database for the system were the closest to a real-time application where users do not have the time or patience for lengthy recordings.

The database was recorded and processed by the team and it is composed of 30 speakers (24 men and 6 women) with ages from 20 to 30 years. The phrases are composed of the first and last name of each speaker spoken in Portuguese, with 10 true phrases and 5 false phrases. The recording was performed in a quiet environment, with fixed phone. The recorded material was stored in wav mono, 16 bits. Table 1 summarizes the information from database.

Table 1: Speakers Data Base Description.

Speakers	30 (24 males/6 females)
Section/Speaker	1
Content	Name and Last name
Channel	Phone or mobile-phone
Acoustic Environment	Various
Idiom	Portuguese
Audio Sample Size	16 bits
Audio Sample Rate	16 KHz
Format file	Wave

4.2 Feature Extraction

The feature extraction process began applying a voice activity detection algorithm used to discard silence–noise frames. The speech activity detector is a self-normalizing, energy based detector that tracks the noise floor of the signal and can adapt to changing noise conditions.

Next, the speech was segmented into frames by a 20-ms window progressing at a 10-ms frame rate and a mel-scale cepstral feature vectors was extracted from the speech frames. The mel-scale cepstrum is the discrete cosine transform of the logspectral energies of the speech segment.

The features consists of the widely used 12 mel-frequency cepstral coefficients (MFCCs) using C0 as the energy component, appended with delta and acceleration coefficients, and computed every 10 milliseconds (i.e., 10 ms is the frame shift) for a frame of 20 ms.

4.3 Threashold

Speaker verification means making a decision on whether to accept or to reject a speaker. To decide, a threshold is used with each client speaker. If the unknown speaker's maximum probability score exceeds this threshold, then the unknown speaker is verified to be the client speaker (i.e., speaker accepted). However, if the unknown speaker's maximum probability score is lower than this threshold, then the unknown speaker is rejected. The relationship is shown in Figure 2.

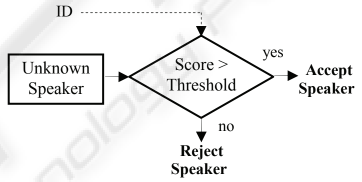

Figure 2: Speaker Verification System.

The threshold is determined as follows:

1. For each speaker, evaluate all samples spoken by him using his own GMM models and find the probability scores. From the scores, find the mean, μ_1, and standard deviation, σ_1 , of the distribution.
2. For each speaker, evaluate all samples spoken by a large number of impostors (typically over 20) using the speaker's HMM models and find the probability scores. From the scores, find the mean μ_2 and standard deviation σ_2 of the distribution.
3. For each speaker, calculate the threshold as Eq. 2:

$$T = \frac{\mu_1\sigma_2 + \mu 2_2\sigma_1}{\sigma_1 + \sigma_2} \quad (2)$$

4.4 Classification

The process of classification began with the division of the database in two different bases for training and testing. The choice of training vectors was made randomly among the 10 training phrases. There was a case of 20 iterations in which 5 phrases were chosen randomly for training and the other 5 for testing. This ensured good variability and consistency of the model reference.

Then the next step was calculating the threshold. As the reference models were generated for each speaker they were tested on a validation for calculating the average distance between true speakers and impostors. This measurement is stored to serve as the calculation of decision of the stage of testing.

The tests database was divided into two groups. With a genuine phrases (the 5 other phrases of the stage of training) and another with the phrases of impostors (5 separate impostors). The process of testing was also performed in 20 interactions for 16, 32 and 64 mixtures. For both groups were estimated parameters of Maximum Likelihood.

The result of these distortions was compared to the threshold calculated in training step and the basic measures of error used in the system were the False Acceptance Rate (FAR) and False Rejection Rate (FRR) as defined below.

$$\text{FAR} = \frac{\text{Number of accepted imposter claims}}{\text{Total number of imposter accesses}} \text{ x } 100$$

$$\text{FRR} = \frac{\text{Number of rejected genuine claims}}{\text{Total number of genuine accesses}} \text{ x } 100$$

Overall performance can be obtained by combining these two errors into total success rate (TSR) where:

$$\text{TSR} = 100\% - \left(\frac{\text{FAR} + \text{FRR}}{\text{Total number of accesses}} \right) \text{ x } 100$$

4.5 Results

Table 2 shows a summary of the results of the use of speaker verification. Have seen the results for 16, 32 and 64 and the results were mixed with 64 blends the best as expected. The overall success rate (TSR) was up 91.12%. For the false acceptance rate (FAR) and false rejection (FRR), the results were very similar, with no great increase in performance.

Table 2: Speakers Data Base Description.

Mixtures	FAR	FRR	TSR
16	11,23	8,92	89,92%
32	10,98	8,47	90,27%
64	9,43	8,32	91,12%

5 CONCLUSIONS

This article show a speaker verification system based on the GMM approach to telephony environments. This system was integrated with the Asterisk telephony platform and the results are very similar to systems for quiet environments.

The tests using the GMM approach results in the range of 90%. Other techniques being worked see better results in the literature, however, our idea was to show the ease of integration between the platform and the asterisk of speaker verification system.

As future work we test in more noisy acoustic environments with cell phones, make changes in the extraction of features to achieve more representative parameters and finally implement the changes as the GMM as FGMM (Tran, 1998) and Type-2-Fuzzy GMM - (Zeng, 2007). Thus we want to achieve better recognition rates and robustness to noisy environments.

REFERENCES

Naik, J.M., Netsch, L.P., Doddington, G.R. 1989. *In International Conference on Acoustics, Speech, and Signal Processing*. Speaker verification over long distance telephone lines.

Mak, M.W., Kung, S.Y. 2002. *In International Conference on Acoustics, Speech, and Signal Processing*. Combining stochastic feature transformation and handset Identification for telephone-based speaker verification.

Meggelen, J. V., Madsen, L., Smith, J., 2007. *Asterisk – The Future of Telephony,* O'REILLY. USA, 2nd edition.

Spencer, M., et al., 2003. *The Asterisk Handbook*, Digium, Inc. USA, 2nd version.

Campbell, J.P. 1997. *In Proceedings of IEEE*, Speaker Recognition: A Tutorial.

Mashao, D., Skosan, M. 2006. *In Pattern Recognition,* Combining Classifier Decision for Robust Speaker Identification.

Fechine, J.M. 2002. *In PhD Thesis, UFCG-Brazil*. Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística.

Reynolds D., R. Rose. 1995. *In IEEE Trans. Speech Audio Process.* Robust text-independent speaker identification using gaussian mixture speaker models.

Reynolds, D., T. Quatieri, R. Dunn. 2000. *In Digital Signal Process., DSP*. Speaker verification using adapted Gaussian speaker mixture models.

Tran, D., VanLe, T., Wagner M., 1998. *In: Proceedings of the International Conference on Spoken Language Processing.* Fuzzy Gaussian mixture models for speaker recognition.

Zeng, J., Xie, L., Liu, Z.Q., 2007. *In Pattern Recognition,* Type-2-Fuzzy Gaussian Mixture Models.