*ELECTRONICS AND ELECTRICAL ENGINEERING*

*ISSN 1392 – 1215* *2008. No. 5(85)*

*ELEKTRONIKA IR ELEKTROTECHNIKA*

*SYSTEM ENGINEERING, COMPUTER TECHNOLOGY*

*T 120*

*SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS*

# Speaker Recognition using Gaussian Mixture Models

**J. Kamarauskas**

*Institute of Mathematics and Informatics,*
*A. Goštauto str. 12, LT–01108 Vilnius, Lithuania; e-mail: j.kamarauskas@ltec.lt*

## Introduction

Speech signal contains several levels of information. At first it contains information about the spoken message. At second level speech signal also gives information about the speaker identity, his emotional state and so on.

The task of speaker recognition can be divided into two parts: speaker identification and speaker verification. Speaker identification is answering the question which one of the group of known voices best matches the input voice. Speaker verification is answering the question is really this person who claims to be. Also speaker recognition can be text dependent or text independent. In text dependent speaker recognition, speech recognition is performed too and there are used the same methods as in speech recognition.

In speech and speaker recognition systems various features are used [1], calculated from the short intervals (named as frames) of the speech signal: coefficients of Linear prediction coding (LPC), cepstral coefficients, calculated from LPC model (LPCC), mel-cepstrum coefficients (MFCC), bark cepstrum coefficients, delta cepstrum and so on. Duration of the frame is about 25ms. These frames overlap one another. The same features are often used in speech and speaker recognition systems, however there are two completely different tasks.

There are proposed a lot of methods for speaker modelling and recognition. In text dependent speaker recognition the most popular methods are dynamic time warping (DTW), Hidden Markov Models (HMM) [2]. In text independent speaker recognition the most popular methods are: Vector Quantization (VQ) [3], fully connected (ergodic) HMM‘s, artificial neural networks (ANN) [4], support vector machines (SVM) [5], and Gaussian Mixture Models (GMM) [6].

In this paper we would like to propose text independent speaker recognition method with new feature vectors, that consist of fundamental frequency and four formant frequencies, try to build Gaussian Mixture speaker models. Vector Quantization method was employed for initial parameters estimation of speakers GMM. Experiments of speaker recognition were performed and compared with experiments using Gaussian Mixture Models with mel – frequency cepstral coefficients, that is baseline in speaker recognition.

## Features of the speech signal

Now we consider some features of the speech signal, which were used in this research.

Speech production can be divided into three stages [7]: first stage is the source production, second stage is the articulation by vocal tract, and the third stage is the sound propagation from the lips and nostrils. The most important components of human speech production system are the lungs, trachea, larynx, nasal cavity velum, hard palete tongue, teeth and lips. All these components are called articulators. They move to the different positions to produce various sounds. There are three main cavities: nasal oral and pharyngeal that comprise the main acoustic filter. Combination of these cavities and articulators is called vocal tract. A voiced sound is generated by vibratory motion of the vocal cords powered by airflow generated by expiration. The frequency of oscilation of vocal cords is called the fundamental frequency. Unvoiced sound is produced by turbulent airflow passing through a narrow constriction in the vocal tract.

Modeling process is divided into two parts: the excitation (source) modeling and the vocal tract modeling [8]. This approach is based on the assumption of independence of the source and the vocal tract models.

Voiced signals can be modeled as a fundamental frequency signal filtered by the vocal tract and unvoiced is a white noise also filtered by the vocal tract. We can think about vocal tract as a digital filter that affects source signal, so from the filters output we can extract filter parameters.

In the model of linear prediction (LPC) coding, the speech signal is shown as an autoregression sequence. It is considered that in short time intervals the vocal tract is time-invariant, therefore the value of the signal can be approximately predicted having a certain count of the previous signal values in their linear combination

$$\hat{s}[n] = \sum_{i=1}^{p} a_i \hat{s}[n-i] + Gu[n], \qquad (1)$$

where $\hat{s}[n]$ is the nth predicted value of the signal, u[n] is an error signal, G is the amplification coefficient which makes the energy of the actual and the predicted signal equal, p is the order of the LPC model.

Coefficients of the predicted filter $a_i$ are calculated by minimazing energy of the error signal. For this purpose the Durbin algorithm is often used [9].

Fundamental frequency can be calculated from the excitation signal that can be calculated having parameters of LPC model and can be expressed as:

$$u[n] = s(n) - \sum_{i=1}^{p} a_i \hat{s}[n-i]. \quad (2)$$

After that low-pass filter with cut-off frequency at 3000Hz is used and autocorelation function of the residual (excitation) signal is calculated. The distance between two peaks of the autocorelation function corresponds to the fundamental frequency.

If we look at the Fourier spectrum of the signal frame we will see there some peaks, what are called formants. In the frequency range 200-5000Hz we can see 3-5 maximas. Each formant corresponds to a resonance in the vocal tract. Positions of the formants are well seen if we look at transfer function of the vocal tract. We can calculate transfer function from the LPC parameters, that corresponds to the vocal tract.

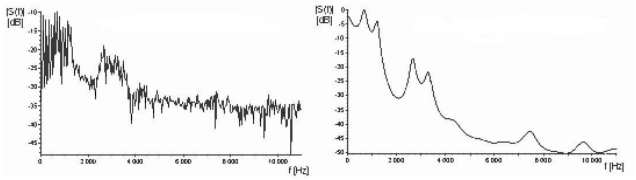

**Fig. 1.** Fourier transform of signal frame and transfer function calculated from the LPC parameters

In the left side of Fig. 1 Fourier transform of the signal frame of the vowel a is shown. In the right side transfer function calculated from the LPC parameters of this frame is shown, where positions of the formants are seen visibly.

Calculation of the formants is the task complicated enought. This is because maximas of the spectrum disappear in certain conditions and their calculation from the envelope of the spectrum becomes imposible. Method of the line spectral pairs [10] was used for this purpose.

**Gaussian Mixture Models**

Gaussian Mixture density is weighted sum of M component densities and can be expressed:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}), \quad (3)$$

where $\vec{x}$ is D dimensional vector, $p_i$ is the component weight, $b_i(\vec{x})$ - component densities, that can be written:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \left| \sum_i \right|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' \sum_i^{-1} (\vec{x}-\vec{\mu}_i)}, \quad (4)$$

where $\mu_i$ - mean vector, $\sum_i$ - covariance matrix.

Mixture weights must satisfy constraint:

$$\sum_{i=1}^{M} p_i = 1. \quad (5)$$

Gaussian mixture density is parametrized by the mean vectors, covariance matrices and mixture weights. All these parameters are represented by notation:

$$\lambda = \left\{ p_i, \mu_i, \sum_i \right\} \; i = 1,2,...,M. \quad (6)$$

So, each speaker is represented by his/her GMM and is referred by his/her model λ

The other task is to estimate the parameters of GMM λ, which best matches the distribution of the training feature vectors, given by speech of the speaker. There are several available techniques for GMM parameters estimation [11]. The most popular method is maximum likehood (ML) estimation [12]. The basic idea of this method is to find model parameters which maximaze the likehood of GMM. For a given set of T training vectors $X = \{\vec{x}_1,...,\vec{x}_T\}$ GMM likehood can be written:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda). \quad (7)$$

ML parameter estimates can by obtained iteratively using special case of expectation-maximization (EM) algorithm. There the basic idea is, beginning with initial model $\lambda$, to estimate a new model $\overline{\lambda}$, that $p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$. The new model then becomes the initial model for the next iteration. This process is repeated until some convergence threshold is reached.

On each iteration, folowing reestimation formulas are used: mixture weights are recalculated

$$\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda). \quad (8)$$

Means are recalculated

$$\overline{\vec{\mu}}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)}. \quad (9)$$

Variances are recalculated

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)(x_t - \mu_i)^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)}. \quad (10)$$

The a posteriori probability for acoustic class i is given by:

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)}. \quad (11)$$

Vector quantization algorithm was used for finding initial parameters of GMM. Count of clusters was equal to the count of Gaussian mixtures. Means and variances of clusters were taken as initial parameters of GMM. Initial mixture weights were calculated as ratio of feature vectors, that makes a cluster with all feature vectors.

## Experimental results

Experiments of speaker recognition were performed using three different features: four formants, four formants and fundamental frequency, and mel-frequency cepstrum coefficients (MFCC) of 13-teen order. Speech database consisted of forty two male speakers, where every person had pronounced the same phrase fifteen times. Three or four phrases were used to build Gaussian mixture speaker models for every speaker and every type of features and other phrases were used for testing. Count of mixtures differed from 12 to 16.

In Fig. 2 curves of intraindividual and interindividual distortions of the likehood are shown using mel frequency cepstral coefficients.

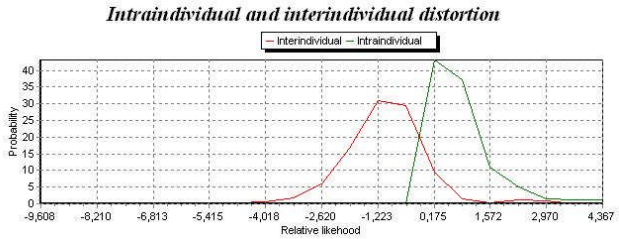


**Fig. 2**. Intraindividual and interindividual distortion using MFCC

Intraindividual distortions are obtained when phrases and models of the same speaker are compared. Interindividual distortions are obtained when phrases and models of the different speakers are compared. In the ideal case there is no intersection between intraindividual and interindividual distortions and this intersection leads to the recognition mistakes, that can be of the two types: false accept (FA) and false reject (FR). False accept error is when impostor is accepted as own person. False reject error is when the own person is discarded as impostor.

In the Fig. 3 FAR – FRR and DET (Detect Error tradeoff) curves are shown using MFCC. These curves are often used to represent accuracy of recognition results.

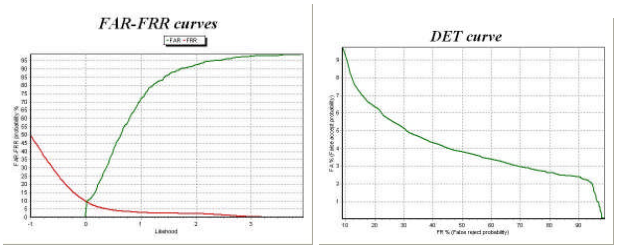


**Fig. 3**. FAR-FRR and DET curves using MFCC

The equal error rate (ERR) 9.6% was obtained.

In Fig. 4 intraindividual and interindividual distortions are shown using four formants frequencies and in the Fig. 5 corresponding FAR-FRR and DET curves are shown.

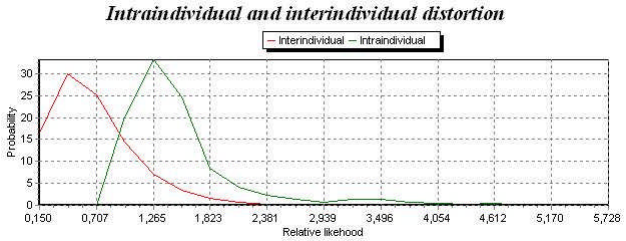


**Fig. 4**. Intraindividual and interindividual distortion using formant frequencies

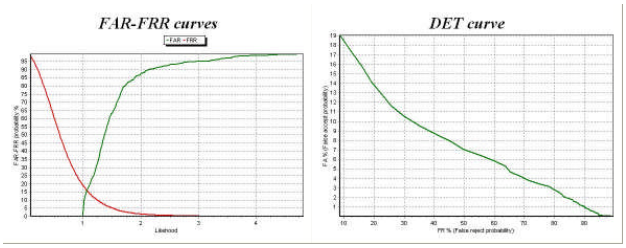


**Fig. 5**. FAR-FRR and DET curves using formant frequencies

The equal error rate (ERR) 15.1% was obtained.

In Fig. 6 intraindividual and interindividual distortions are shown using four formants frequencies with fundamental frequency and in the Fig. 7 corresponding FAR-FRR and DET curves are shown.

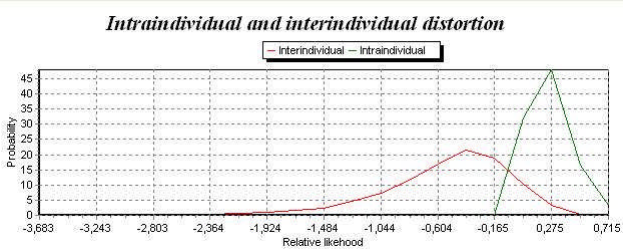


**Fig. 6**. Intraindividual and interindividual distortions using formant frequencies and fundamental frequency

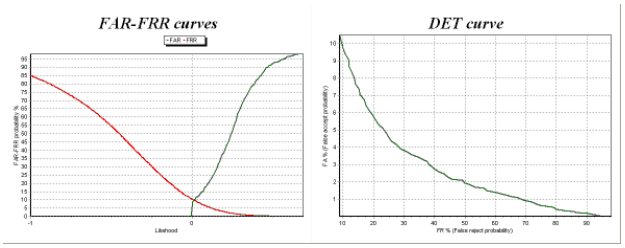


**Fig. 7**. FAR-FRR and DET curves using formant frequencies and fundamental frequency

The equal error rate (ERR) 9.7% was obtained in this case.

## Conclusions

1. Formant frequencies and fundamental frequency can be used as a features for speaker recognition task.
2. Results of accuracy of speaker recognition achieved using mel – frequency cepstral coefficients and formant frequencies with fundamental frequency are quite the same. But parameters estimation of GMM using mel – frequency cepstral coefficients continues few times longer.

3. The worst results were achieved using as the features formant frequencies only.

**References**


1. **Picone J.** Signal Modeling Techniques in Speech Recognition // Proceedings of the IEEE. – 1993.
2. **Rabiner L. and Juang B.H.** An introduction to hidden Markov models // IEEE ASSP Mag. – 1986. – P. 4–16.
3. **Juang B.H., Wang D. Y. and Gray A. H.** Distortion Performance of Vector Quantization for LPC Voice Coding. // IEEE Tranc. on Acoustic Speech and Signal Processing. – 1982. – Vol. ASSP-30, No. 2. P. 294–304.
4. **Navakauskas D.** Skaitmeninio signalų apdorojimo priemonės. Dirbtinių neuronų tinklai. – Vilnius: Technika, 2000.
5. **Vapnik V. N.** Statistical Learning Theory. – New York: Wiley, 1998.
6. **Reynolds D. A. and Rose R. C.** Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models // IEEE Transactions on Speech and Audio Processing. – January 1995. – Vol. 3, No. 1. – P. 72–83.
7. **Furui S.** Digital Speech Processing, Synthesis and Recognition. – New York: Marcel Dekker, 2001.
8. **Deller J. R., Hansen J. H. L., Proakis J. G.** Discrete-Time Processing of Speech Signals. – Piscataway (N.J.): IEEE Press, 2000.
9. **Rabiner L. and Juang B. H.** Fundamentals of Speech Recognition. – Prentice-Hall, Englewood Cliffs, New Jersey, USA. – 1993.
10. **Kabal P. and Ramachandran R. P.** The Computation of Line Spectral Frequencies Using Chebyshev Polynomials. // IEEE Transactions on Acoustic, Speech, and Signal Processing. – December 1986. – Vol. ASSP-34, No. 6. – P. 1419–1426.
11. **McLachlan G.** Mixture Models. – New York: Marcel Dekker, 1988.
12. **Dempster A., Laird N., and Rubin D.** Maximum likelihood from incomplete data via the EM algorithm // J. Royal Srar. Soc. – 1977. – Vol. 39. – P. 1–38.


Submitted for publication 2008 02 15


**J. Kamarauskas. Speaker Recognition using Gaussian Mixture Models // Electronics and Electrical Engineering. – Kaunas: Technologija, 2008. – No. 5(85). – P. 29–32.**

Gaussian Mixture models is one of the most popular statistical methods in speaker recognition. The purpose of this research is to perform experiments of speaker recognition using various feature vectors: four formants, four formants with fundamental frequency and mel cepstrum coefficients. Gaussian mixture models using mel cepstrum coefficients is baseline in speaker recognition and gives one of the best results in text independent speaker recognition. After implementing experiments of speaker recognition and comparing experimental results we can affirm that mel scale cepstral coefficients and four formants with fundamental frequency gives quite the same recognition accuracy, but creating of Gaussian mixture speaker models and recognition process continues a few times longer using mel scale cepstral coefficients, because count of calculations is few times greater in that case. Using only four formants gives the worst results of recognition accuracy. Ill. 7, bibl. 12 (in English; summaries in English, Russian and Lithuanian).

**Ю. Камараускас. Распознавание говорящего используя модель гауссовых смесей // Электроника и электротехника. – Каунас: Технология, 2008. – № 5(85) – С. 29–32.**

Модели гауссовых смесей являются одними из самых распостранённых статистических методов разпознавания говорящего. Цель этого исследования – провести эксперименты распознавания говорящего используя разные векторы признаков: четыре форманты, четыре форманты с основным тоном и мел–кепстральные коэффициенты. Модели гауссовых смесей с мел–кепстральными коэффициентами являются одними из самых распостранёных методов, дающих одни из самых лучших результатов в независимом от текста распознавании говорящего. После проведенных экспериментов распознавания говорящего и сравнения экспериментальных резултатов можно делать вывод, что мел–кепстральные коэффициенты дают почти ту же самую точность распознавания, что и четыре форманты с основным тоном, но создание модели гауссовых смесей и процесс распознавания используя мел–кепстральные коэффициенты требуют несколько раз больше времени, чем форманты с основным тоном, потому что в этом случае нужно выполнить несколько раз больше вычислительных операций. Худшие результаты разпознавания получаются используя в место признаков только четыре форманты. Ил. 7, библ. 12 (на английском языке; рефераты на английском, русском и литовском яз.).

**J. Kamarauskas. Kalbančiojo atpažinimas taikant Gauso mišinių modelius // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2008. – Nr. 5(85). – P. 29–32.**

Gauso mišinių modeliai – vienas iš plačiausiai taikomų statistinių kalbančiojo atpažinimo metodų. Šio tyrimo tikslas – atlikti kalbančiojo atpažinimo eksperimentus naudojant įvairius požymių vektorius: keturias formantes, keturias formantes kartu su pagrindiniu tonu bei melų skalės kepstro koeficientus. Gauso mišinių modeliai su melų skalės kepstro koeficientais yra jau tapę klasikiniu asmens atpažinimo metodu, duodančiu vienus iš geriausių nepriklausomo nuo teksto kalbančiojo atpažinimo rezultatų. Atlikus asmens atpažinimo eksperimentus ir palyginus gautus rezultatus galima daryti išvadas, kad tiek melų skalės kepstro koeficientai, tiek keturios formantės kartu su pagrindiniu tonu duoda beveik tą patį atpažinimo tikslumą, tačiau Gauso mišinių modelio kūrimas bei atpažinimas panaudojant melų skalės kepstro koeficientus trunka kelis kartus ilgiau, nes šiuo atveju reikia atlikti kelis kartus daugiau skaičiavimo operacijų. Prasčiausi atpažinimo rezultatai gaunami naudojant tiktai keturias formantes. Il. 7, bibl. 12 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).