ELECTRONICS AND ELECTRICAL ENGINEERING

ISSN 1392 – 1215 2010. No. 2(98)

ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY

T 120

SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

Evaluation of Effectiveness of Different Methods in Speaker Recognition

B. Šalna

Forensic Science Centre of Lithuania,
Lvovo str. 19a, LT–09313 Vilnius, Lithuania; e-mail: bernardas@centras.lt

J. Kamarauskas

Vilnius Gediminas Technical University,
Saulėtekio av. 11, LT–10223Vilnius, Lithuania; e-mail: juozas.kamarauskas@gmail.com

Introduction

A big attention is paid to biometrics technologies now and material and intelectual resources, testing centres have been established. If the other kinds of biometrics need special devices and corresponding infrastructure must be created (for example for iris), voice biometrics does not need it. Therefore a big attention is paid to creation of algorithms of speaker identification by voice all over the world and according to predictions, solutions are waiting of voice biometrics in criminology (automatic speaker recognition by voice), mobile banking and internet marketing.

In spite of great achievements in speaker recognition technology there is no theory created how does human separate one voice from the other and there is no system of features created that would let separate two voices having different phrases, speaking environment, sound recording channels and so on.

Effectiveness of methods of speaker identification depends on feature system and comparison method. A big attention is paid to the speech recognition now (converting speech signal to text or control by voice) and systems of features that are created for this purpose are used for speaker recognition too. These features are considered contrarily in literature. One states that these features do not depend on speaker individuality and used in speech recognition applications [1, 2], others state that they represent speaker individuality very well [3, 4]. It is worth mentioning that appears more and more works where new systems of features, used for speaker recognition are creating and analyzing now [5, 6].

Voice biometrics gives worse results compared to other kinds of biometrics but it could be widely used. Therefore investigations in that field should be made.

We would like to propose our method for speaker recognition named as phonemic method and compare this method with baseline method and others that are often used in speaker recognition.

Speaker recognition systems

An algorithms of speaker recognition can be divided into two groups:

- Text-dependent speaker recognition;
- Text-independent speaker recognition.

Procedure of speaker recognition consists of two stages:

1. Training stage. During this stage feature vectors are calculated from the speech signal. A speaker model or VOICEPRINT is built using these feature vectors.
2. Recognition or verification stage. Feature vectors are calculated from the uttered phrase of speech signal of unknown speaker. These features are compared against speaker model or VOICEPRINT. The obtained similarity score (or difference) is compared against threshold, set to this speaker. After that decision about speakers identity is made.

The ROC and DET curves are used for evaluation of effectiveness of voice biometric systems.

The UBM-GMM model is baseline in speaker identification or voice biometric systems now. During analysis the speech signal is divided into segments of equal length – frames (about 20-25 ms). These frames overlap one another. A feature vector is calculated from the signal frame as shown in Fig. 1.

The standard feature vector consists of 39 elements:

- 12 MFCC,
- 12 delta- MFCC (ΔMFCC),
- 12 delta-delta MFCC (ΔΔMFCC),
- 1 (log) frame energy,
- 1 (log) delta frame energy,
- 1 (log) delta-delta frame energy.

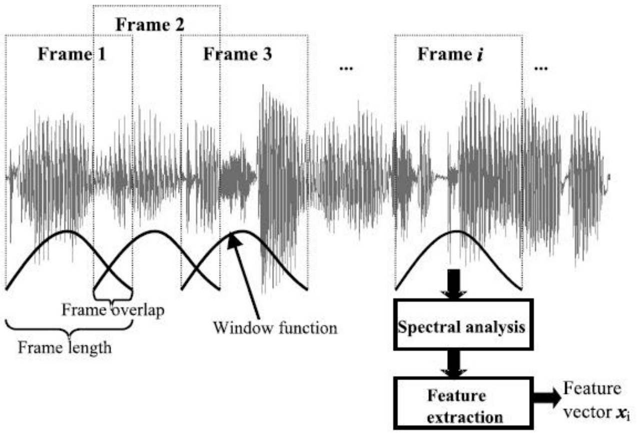

Fig. 1. Feature extraction from the speech signal [7]

There are proposed a lot of methods for speaker modeling and matching. In text-independent speaker recognition the most popular methods are:

- Gaussian mixture models (GMM) [8];
- Vector quantization (VQ) [9];
- Artificial neural networks (ANN) [10];
- Support vector machines (SVM) [11];
- Fully-ergodic hidden Markov models (HMM) [12].

In text-dependent speaker recognition the same methods as in speech recognition are used. The most popular are:

- Dynamic time warping (DTW) [13];
- Hidden Markov models (HMM) [12].

Phonemic speaker recognition method

Our proposed system of features and comparison method we named as phonemic method. Feature system consists of 36 components that represents individual features of speaker. This system of features consists of four formant frequencies, two antiformant frequencies, 4 normalized amplitudes of formants and other combination parameters of spectral pairs.

If we look at the Fourier spectrum of the signal frame we will see there some peaks, what are called formants and valleys, called antiformants. In the frequency range 200-5000Hz we can see 3-5 maximums. Each formant corresponds to a resonance in the vocal tract.

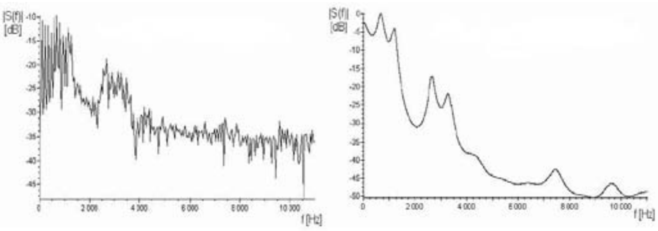

Fig. 2. Fourier transform of signal frame and transfer function calculated from the LPC parameters [6]

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.

In the left side of Fig. 2 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 enough. This is because maximums of the spectrum disappear in certain conditions and their calculation from the envelope of the spectrum becomes impossible. Method of the line spectral pairs [14] was used for this purpose.

Lets denote F(i) – frequency of i-th formant, ANF(i) – frequency of i-th antiformant, LSF(i) - frequency value of i-th spectral pair. If the sampling rate of the speech signal is 11025Hz and order of LPC model is 12, we propose next experimential formulas for formant and antiformant estimation [15]:

$$\begin{cases} F(1) = LSF(2), \\ F(2) = LSF(5), \\ F(3) = LSF(8), \\ F(4) = LSF(11), \\ ANF(1) = (LSF(2)+LSF(3))/2, \\ ANF(2) = (LSF(5)+LSF(6))/2, \\ ANF(3) = (LSF(8)+LSF(9))/2. \end{cases} \quad (1)$$

Spectrogram of the signal of phonemes A E and I and positions of linear spectral pairs is shown in Fig. 3.

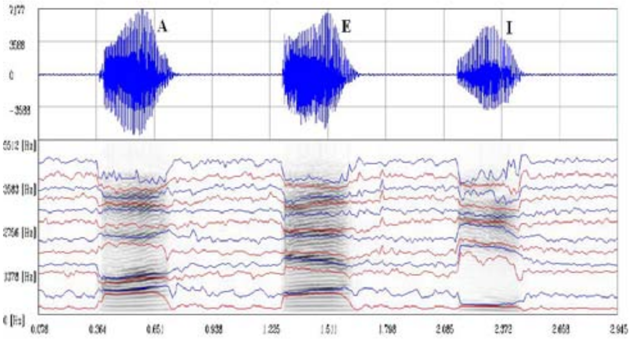

Fig. 3. Speech signal, spectrogramm and linear spectral pairs

There are next theoretical and practical motivations for using phonemic method in speaker recognition. As we can see in Fig. 4, distance between two different phonemes /A/ and /I/ of the same speaker in the spectral domain is always bigger than distance between the same phoneme /I/ of the different speakers. According to formant speech generation theory the first 2-3 formants make the biggest influence in forming different sounds or phonemes, formants of higher order reflect speaker individuality. Therefore we should select the same type of phonemes or acoustic events from the speech signal during speaker recognition, calculating shape of vocal tract in every frame and comparing feature vectors of frames with similar shape of the vocal tract.

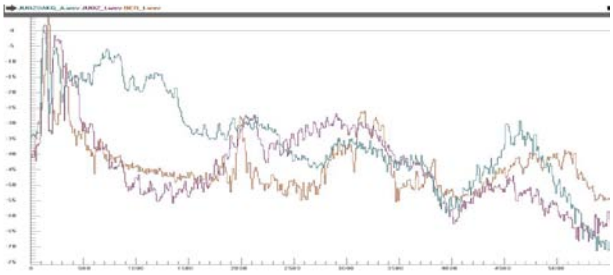

Fig. 4. Spectrums of different phonemes of different speakers

We will consider phonemic method below.

Feature matrix is calculted for every utterance during training (three phrases often are used to build speaker‘s model). This matrix consists of feature vectors, that are calculated from the voiced speech segments – frames. Dimension of feature matrix is N×36, where N is count of voiced frames. The voice template of the speaker (speaker‘s model or voiceprint) is calculated from these three feature matrixes. Steps of calculating VOICEPRINT are described next. Comparison of two feature matrixes is performed first. Vocal tract shape from the first three formants of the first frame are calculated and vocal tract shape is calculated for the first frame of the second feature matrix in the same way too and difference between these two shapes is calculating. Then the vocal tract shape of the second frame of second feature matrix is calculating and difference between it and vocal tract shape of the first frame first feature matrix is found. After that third feature vector of second feature matrix is taken for calculating vocal tract shape and so on until we find the nearest feature vector in the second feature matrix to the first feature vector of first feature matrix. When the nearest vector is found, absolute distance of all alements of two feature vectors is calculated. Thus voiceprint consist of one reference matrix and variation bounds of elements of this matrix. Now we calculate variation of these differencies using histograms. After that shape and other statistical parameters of these histograms are evaluated and common statistical distance is obtained. This distance corresponds to variation bounds of speaker or intraindividual distance.

During verification process feature matrix is calculated from the uttered phrase and it is compared against template matrix in the same way like in training process. If uttered phrase is compared agains the template (model) of the same speakers, obtained realative distance is in the range, obtained during training phase and is small. In the case of different speakers it is big.

Phonemic method has been used in our created software packet SIVE, it consists of several submodules. This method can be text-dependent or text-independent. Text-dependent version designed to speaker verification we named BALSO RAKTAS (RAKTAS).

Experimental results

Four recognition systems were used for speaker recognition experiments:

1. Speaker recognition system that used LPCC of 22 order as features and vector quantization [15] (VQ) approach for the pattern classification was used. Count of centroids was equal to 32.
2. Speaker recognition system that uses MFCC of the 13 order as features and Gaussian mixture models (GMM) for speaker modelling and pattern clasification. Count of mixture components was equal to 16. This system is baseline in speaker recognition.
3. Speaker recognition system that uses 4 formants three antiformants and pitch (F0) as features and Gaussian mixture models (GMM) for speaker modelling and pattern clasification. Count of mixture components was equal to 20.
4. Proposed speaker recognition system, using phonemic method.

We have implemented speaker recognition experiments using two speech databases

1. Russian Speech database (RUSBASE) which is distributed by ELRA (European Language Resources Association). We used first phrase from this database. It is repeated in 3 sessions 5 times. Recordings were made in the lab using good microphones. Voices of 41 men were used only.
2. Database, created by firm “PORTICUS” from USA. There are 39 speakers (24 women and 15 men). Phrases were uttered using mobile phone during four sessions in different places. Uttered phrase – sequence of digits “8-7-2-3-1-5-9-4-6-0”.

Experimental results of the speaker verification using VQ GMM and RAKTAS and the same speech databases are given in the table (Equal error rate – EER is given).

Table 1. Experimental results of speaker recognition using different methods

Method	RUSBASE (EER,%)	PORTICUS P4 (EER,%)
VQ-LPCC	2,76	6,64
GMM-MFCC	5,86	6,81
GMM-4F3AF0	5,17	7,28
RAKTAS	2,32	10,65

Conclusions

1. Proposed method for speaker recognition based on formant and antiformant features, calculated from linear spectral pairs and their combination parameters.
2. Experiments of speaker recognition were performed using two databases and different methods.
3. Proposed (phonemic) method outperformed baseline methods used for speaker recognition – GMM with MFCC.

References

1. **Umbach R., Loog M.** An investigation of Cepstral Parametrisations for Large Vocabulary Speech Recognition // Eurospeech 99, 6–th European Conference on Speech Communication and Technology. – Budapest, Hungary. – 1999. – P. 1323-1326.
2. **Vergin R., O‘Shaughnessy D., Farhat A.** Generalized Mel Frequency Cepstral Coefficients for Large – Vocabulary Speaker – Independent Continuous – Speech Recognition //

IEEE Trans. On Speech and Audio Processing. – 1999. – No. 7(5). – P. 525–532.

3. **Miuller Ch**. Speaker Classification. – Berlin Heidelberg: Springer–Verlag, 2007.
4. **Scheffer N, Bonastre J. F.** A UBM-GMM driven discriminative approach for speaker verification // In Proc. ODYSSEY. – 2006. – P. 1–7.
5. **Gudnason J**. Voice Source Cepstrum processing for Speaker Identification / PhD Thesis. – Imperial College London, University of London. – 2007.
6. **Orsag F**. Biometric Security Systems, Speaker Recognition Technology / Dissertation, BRNO university of technology. – 2004.
7. **Kinnunen T. H**. Spectral Features for Automatic Text-Independent Speaker Recognition / Licentiate‘s Thesis, University of Joensuu. –2003.
8. **Reynolds D., Rose R**. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models // IEEE transactions on speech and audio processing. – 1995. – No. 3(1). – P. 72–83.
9. **Juang B.-H., et al.** A vector quantization approach to speaker recognition // AT & T Technical Journal. – 1987. – No. 66. – P. 14–26.
10. **Simpson P. K**. Artificial Neural Systems: Foundations, Paradigms, Applications, And Implementations. – New York: Pergamon Press, 1990.
11. **Vapnick V**. The Nature of Statistical learning theory. – New York: Springer-Verlag, 1995.
12. **Rabiner L, Juang B. H**. An introduction to hidden Markov models // IEEE ASSP Mag. – 1986. – No. 3(1). – P. 4–16.
13. **Rabiner R. L., Rosenberg A. E., Levinson S. E**. Considerations in dynamic time warping algorithms for discrete word recognition // IEEE Transactions on Acoustics, Speech and Signal Processing. – 1978. – No. 26(5). – P. 575–582.
14. **Kabal P., Ramachandran R. P**. The Computation of Line Spectral Frequencies Using Chebyshev Polynomials // IEEE Transactions on Acoustic, Speech, and Signal Processing. – 1986. – ASSP-34(6). – P. 1419–1426.
15. **Salna B., Mambro G. D**. Method and System for Biometric Voice Print Authentication. – Patent USA, No. EP-0001915294. – 2006.

Received 2009 10 27

B. Šalna, J. Kamarauskas. Evaluation of Effectiveness of Different Methods in Speaker Recognition // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 2(98). – P. 67–70.

Speaker's identification by voice is a type of biometric systems. Currently it is recent and rapidly developing science and technology with a lot of areas. Voice biometrics gives one of the worst results compared to other kinds of biometrics. The proposed new speaker's recognition method and a new system of features that consists of 36 components for that purpose. These components are formant and antiformant frequencies, their amplitudes, and various other combination parameters of spectral pairs (ratios of formant, their amplitudes and so on). Experiments with two speech databases showed that the proposed method outperformed the standard methods used for speaker identification - Gaussian mixture models using the mel scale cepstral coefficients (MFCC-GMM) and vector quantization (VQ) method. Ill. 4, bibl. 15 (in English; summaries in English, Russian and Lithuanian).

Б. Шална, Ю. Камараускас. Исследование эффективности методов распознавания говорящего // Электроника и электротехника. – Каунас: Технология, 2010. – № 2(98). – С. 67–70.

Идентификация говорящего является одним из видов биометрических систем. Это недавно появившаяся и быстро развивающаяся область науки и технологий с большим количеством областей приложений. Голосовая биометрия пока что дает одних из самых худших результатов. Предлагается новый метод распознавания говорящего, а для этой цели – и новая система признаков, состоящая из 36 векторных компонентов. Эти компоненты включают частоты формант и антиформант и их и амплитуды, а также различные другие комбинационные параметры спектровых пар (соотношения формант их амплитуд и т. д.). Эксперименты с двумя голосовыми базами показали, что предлагаемый метод превысил стандартные методы, используемые для идентификации личности по голосу – модели гауссовских смесей, используя мел-кепстральных коэффициентов (MFCC-GMM) и метода векторного квантования (VQ). Ил. 4, библ. 15 (на английском языке; рефераты на английском, русском и литовском яз.).

B. Šalna, J. Kamarauskas. Kalbančiojo asmens atpažinimo metodų efektyvumo tyrimas // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2010. – Nr. 2(98). – P. 67–70.

Kalbančiojo asmens atpažinimo metodas priklauso biometrinių sistemų tipui. Tai yra neseniai atsiradusi ir sparčiai besivystanti mokslo ir technologijų sritis, turinti labai daug taikymo sričių. Balso biometrijos rezultatai kol kas vieni iš prasčiausių. Pasiūlytas naujas kalbančiojo asmens atpažinimo metodas bei tam tikslui sukurta nauja kalbos signalų požymių sistema, susidedanti iš 36 požymių vektorių komponenčių. Šias komponentes sudaro formančių bei antiformančių dažniai bei jų amplitudės, taip pat įvairūs kiti kombinaciniai spektrinių porų parametrai (formančių, jų amplitudžių santykiai ir t. t.). Atlikus eksperimentus su dviem balsų bazėmis, paaiškėjo, kad pasiūlytasis metodas pralenkė standartinius metodus, naudojamus asmeniui atpažinti pagal balsą – Gauso mišinių modelius, sukurtus naudojant melų skalės kepstro koeficientus (GMM-MFCC), bei vektorinio kvantavimo (VQ) metodą. Il. 4, bibl. 15 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).