ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 *2011. No. 1(107)*
ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY
T 120
SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

Speaker Recognition using Excitation Source Parameters

J. Kamarauskas
Vilnius Gediminas Technical University,
Saulėtekio av. 11, LT–10223 Vilnius, Lithuania; e-mail: juozas.kamarauskas@gmail.com
B. Šalna
Forensic Science Centre of Lithuania,
Lvovo str. 19a, LT–09313 Vilnius, Lithuania; e-mail: bernardas@centras.lt

Introduction

Estimation of fundamental frequency is a long lasting problem in speech applications such as speech analysis and synthesis, speech coding, speaker recognition, various multimedia applications and so on [1]. However it is difficult to find algorithms that would provide desired accuracy and robustness in bad recording conditions (noise, reverberation).

Pitch corresponds to frequency of vibrating folds during speech generation and it can be used as a parameter of person in biometric systems.

Pitch is often used in speaker recognition as feature. However speaker recognition accuracy using pitch is low compared to features of the vocal tract [2]. But on the other hand pitch is more robust feature to the distortions of the recording channel, different noises and so on than features of the vocal tract [3]. Therefore pitch is often combined with other features of the vocal tract or it can be used alone in such applications as forensic sciences where different recording conditions is the main problem in speaker recognition.

There are proposed a lot of methods for pitch calculation. Broadly all methods can be divided into three groups [4]: time domain methods, frequency domain methods and combined methods. Detectors that calculate correlation function in time domain or frequency domain often are used [5]. Cepstrum is used for pitch evaluation too.

We would like to propose our speaker recognition method by calculating pitch using frequency domain method and Gaussian mixture models (GMM) approach for speaker modeling and recognition.

Pitch calculation algorithm

Speech generation consists of three main stages [6]:

- Sound source production;
- Articulation by vocal tract;
- Sound radiation from lips and/or nostrils.

Voiced sounds are generated by vibratory motion of the vocal cords, powered by airflow generated by expiration. The frequency of oscillation of vocal cords is called as fundamental frequency (F0) or pitch. Unvoiced sounds are produced by turbulent airflow passing through narrow constriction of the vocal tract.

We used frequency domain method for pitch calculation. Algorithm is shown in Fig.1.

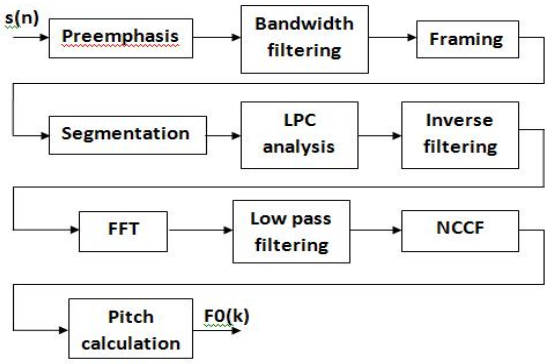

Fig. 1. Pitch calculation algorithm

Preemphasis of speech signal is performed first using 1-st order FIR filter [7]. Its purpose is to amplify components of higher frequencies of spectrum of speech signal.

Then bandwidth filtering is applied using FIR filter of 65 order. To reduce Gibbs effect, filter coefficients are multiplied by Lanczos window. Frequency range depends on speech recording conditions.

Then filtered speech signal is divided in to frames of 30-50 ms length. These frames overlapp one another. Frame shift is equal to 15 ms.

Segmentation is performed next. Its purpose is to remove non-signal frames and background noise. Frames, that do not contain signal are removed first. Sometimes in some recordings there are parts filled by zero or very small

values. Maximum value of the signal is found in the frame and compared against threshold value, equal to 130. If it does not exeed threshold, frame is removed from further calculations.

To find background noise, energy of every frame is calculated. Then 10 frames with minimal energy values are found and energy threshold is calculated from these frames. Frames, that have energy less than threshold are considered as background noise and removed.

LPC analysis is performed using autocorrelation method. Coefficients of the predicted filter a_i are calculated by minimazing energy of the error signal. The Durbin algorithm is used for this purpose [8]. Then inverse filtering is applied to calculate excitation (residual) signal

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

where $u[n]$ – excitation signal; a_i - LPC coefficients; p – LPC order, equal to 8.

Then Fast Fourier transform is applied to the residual signal. We obtain spectrum of the residual signal.

Then normalized cross correlation function (NCCF) of the residual spectrum is calculated [9]

$$NCCF(m) = \frac{\sum_{n=0}^{N-m-1} x(n)x(n+m)}{\sqrt{\sum_{n=0}^{N-m-1} x^2(n) \cdot \sum_{n=0}^{N-m-1} x^2(n+m)}} . \qquad (2)$$

The distance between two peaks of the NCCF corresponds to the fundamental frequency.

To remove some pitch calculation errors derivative of pitch contour is calculated. Pitch can not significantly change in adjacent frames. Pitch values, where derivative of pitch contour has big values, are removed.

Speaker modeling

Histogram techniques are often used to model distribution of the pitch. However distribution of the pitch is not Gaussian. Then differencies or similarities betwen two histograms of comparative records are calculated and decision is made. But comparison results depend on number of classes, used in histograms in this case. We used standard Gaussian Mixture models (GMM) approach for pitch modeling [10, 11]. Thus influence of number of classes is eliminated. Parameter estimation of GMM was done iteratively using special case of the expectation-maximization (EM) algorithm.

In the Fig. 2 real distribution of pitch and approximation by GMM is shown, when 12 classes were used.

As we can see in Fig. 2 – Fig. 4, view of the histograms vary depending on count of classes, so comparison results will vary too. GMM approximation is always the same.

For comparison five measures were used: coincidence, correlation, Euclidean distance, Kullback-Leibler distance and symmetric Kullback-Leibler distance.

Coincidence of two distributions can be calculated

$$Coincidence = \frac{S_{intersection}}{S_{union}} , \qquad (3)$$

where $S_{intersecion}$ is area of intersection between two distributions X and Y and S_{union} is area of union of two distributions.

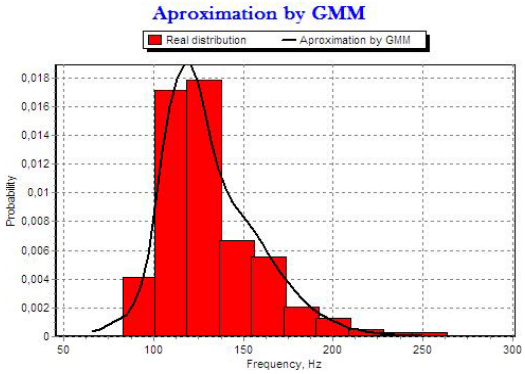

Fig. 2. Real distribution and GMM approximation when 12 classes were used

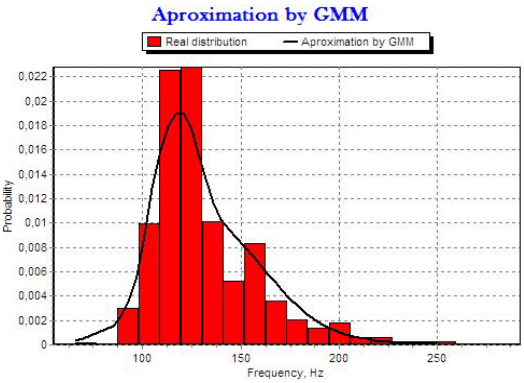

Fig. 3. Real distribution and GMM approximation when 20 classes were used

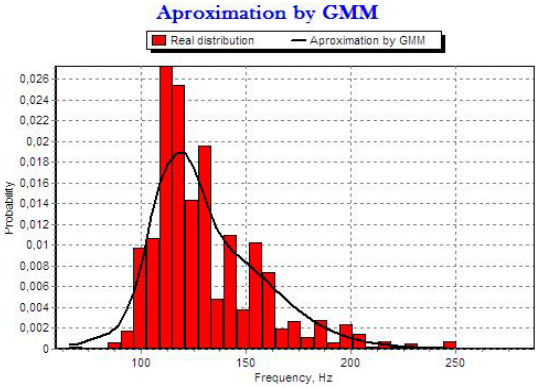

Fig. 4. Real distribution and GMM approximation when 35 classes were used.

Correlation between two distributions X and Y can be calculated

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} , \qquad (4)$$

where S_x and S_y are standard deviations of two distributions X and Y.

Euclidean distance [12] for two distribution X and Y can be calculated

$$d_E(X,Y)=\sqrt{\sum_{i=1}^{N}(x_i-y_i)^2}\ . \qquad (5)$$

Kullback-Leibler (KL) distance [12] (relative entropy) of two distributions X and Y can be expressed

$$d_{KL}(X,Y)=\sum_i x_i\ log\frac{x_i}{y_i}\ . \qquad (6)$$

Because KL distance is not symmetric, so symmetric Kullback-Leibler distance often is used

$$d_{SKL}(X,Y)=\frac{d_{KL}(X,Y)+d_{KL}(Y,X)}{2}. \qquad (7)$$

Experimental results

We have implemented pitch comparison experiments using standard histogram technique and approximation using GMM.

Comparison of histograms of the same speaker are shown in Fig. 5. In Fig. 6 GMM approximations of the same speaker for the same sound recordings are shown.

In the Table 1 there are shown comparison results using histogram techniques and approximations by GMM.

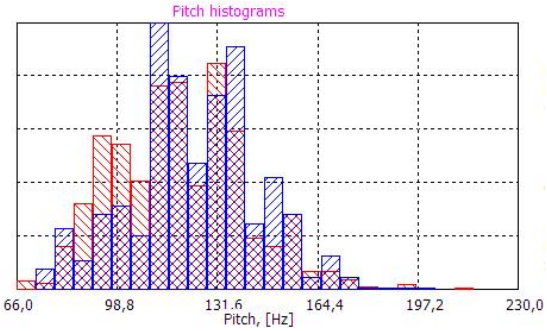

Fig. 5. Comparison of two histograms of the same speaker

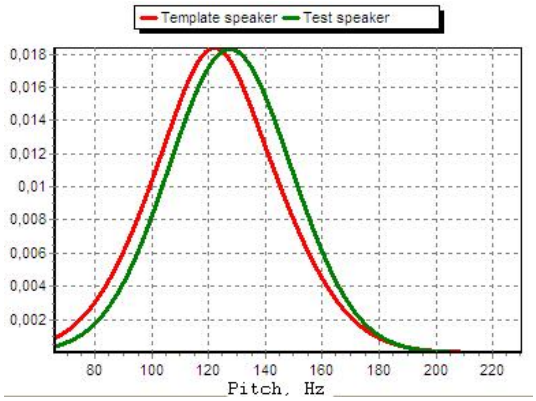

Fig. 6. Comparison of GMM distributions of the same speaker

Table 1. Comparison of pitch distributions of same speaker using different methods

Similarity/Distance	Histograms	GMM
Coincidence	0.69	0.83
Correlation	0.88	0.97
Euclidean dist.	0.11	0.15
Kullback-Leibler	0.21	1.23
Sym. Kullback-Leibler	0.16	1.52

Comparison of histograms of the different speakers are shown in Fig. 7. Fig. 8 shows GMM approximations of the different speakers for the same sound recordings. In the Table 2 there are shown comparison results using histogram techniques and approximations by GMM for the different speakers.

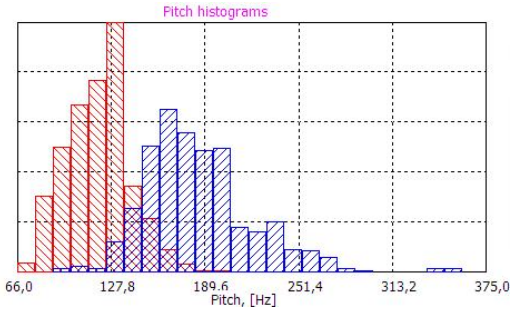

Fig. 7. Comparison of two histograms of the different speakers

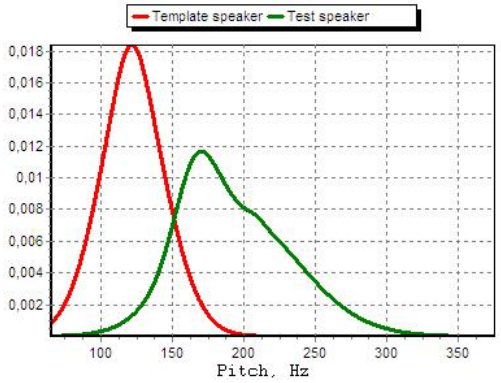

Fig. 8. Comparison of two GMM approximations of the different speakers

Table 2. Comparison of pitch distributions of different speakers using different methods

Similarity/Distance	Histograms	GMM
Coincidence	0.11	0.13
Correlation	-0.13	-0.1
Euclidean dist.	0.46	0.69
Kullback-Leibler	3.9	68.9
Sym. Kullback-Leibler	4.2	89.5

Conclusions

1. Pitch distribution is commonly modeled using histograms. Therefore comparison results depend on count of classes used.
2. By using Gaussian Mixture Models (GMM) for pitch modeling we avoid influence of count of classes.
3. Best comparison results were achieved using symmetric Kullback-Leibler distance (relative entropy).

References

1. **Sharma D., Naylor P. A.** Evaluation of Pitch Estimation in Noisy Speech for Application in non-intrusive Speech Quality Assessment // 17th European Signal Processing Conference (EUSIPCO 2009). – Glasgow, Scotland, August 2009. – P. 2514-2518.
2. **Šalna B., Kamarauskas J.** Evaluation of Effectiveness of different methods in speaker recognition // Electronics and

Electrical Engineering. – Kaunas: Technologija, 2010. – No. 2(98). – P. 67–70.
3. **Kinoshita Y., Ishihara S., Rose P.** Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition // The International Journal of Speech Language and the Law, 2009. – Vol. 16.1. – P. 91–111.
4. **Rabiner L. R., Cheng M. J., Rosenberg A. E., Mcgonegal C. A.** 1976. A Comparative Performance Study of Several Pitch Detection Algorithms // IEEE Transactions on Acoustics, Speech and Signal Processing, 1976. – Vol. ASSP-24(5). – P. 399–418.
5. **Nakasone H., Mimikopoulos M., Beck S. D., Mathur S.** Pitch Synchronized Speech Processing (PSSP) for Speaker Recognition // In Proc. ODYSSEY04. – 2004. – P. 251–256.
6. **Karpov E.** Real time speaker identification. – Masters Thesis. – University of Joensuu, 2003.
7. **Orsag F.** Biometric Security Systems, Speaker Recognition Technology. – Dissertation. – BRNO University of Technology, 2004.
8. **Juang B.-H., et al.** A vector quantization approach to speaker recognition // AT & T Technical Journal, 1987. – No. 66. – P. 14–26.
9. **Kasi K.** Yet Another Algorithm for Pitch Tracking (YAAPT). – Masters Thesis. – Old Dominion University, 2002.
10. **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.
11. **Kamarauskas J.** Speaker recognition Using Gaussian Mixture Models // Electronics and Electrical Engineering. – Kaunas: Technologija, 2008. – No. 5(85). – P. 29–32.
12. **Hautamaki R. E. G.** Fundamental Frequency Estimation and Modeling for Speaker Recognition. – Master‘s Thesis. – University of Joensuu, 2005.

Received 2010 09 23

J. Kamarauskas, B. Šalna. Speaker Recognition using Excitation Source Parameters // Electronics and Electrical Engineering. – Kaunas: Technologija, 2011. – No. 1(107). – P. 55–58.

Excitation signal is used in speaker recognition. It corresponds to the frequency of oscillation of vocal cords and is one of the speaker's characteristics. Although this feature gives worse recognition results compared to the vocal tract parameters, but it is more robust to various distortions in the recording channels. As a result, pitch is commonly used in forensic investigations, where different recording channels is one of the main problems. Currently, the pitch distribution generally is modeled using histograms and calculating various distances or similarity measures between two histograms. However, pitch distribution is not Gaussian and view of the histograms and comparison results depend on the number of classes used. We model pitch distribution using Gaussian mixture models (GMM), and calculate similarity and distance measures between the GMM approximations of two comparative records. Best results were achieved using symmetric Kullback-Leibler distance. Ill. 8, bibl. 12, tabl. 2 (in English; abstracts in English and Lithuanian).

J. Kamarauskas, B. Šalna. Kalbančiojo atpažinimas naudojant žadinimo signalo parametrus // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2011. – Nr. 1(107). – P. 55–58.

Žadinimo signalas naudojamas kalbančiajam atpažinti. Jis atitinka balso stygų virpėjimo dažnį, ir tai yra viena iš kalbančiojo charakteristikų. Nors atpažinimo pagal tokį požymį rezultatai būna prastesni nei pagal balso trakto parametrus, tačiau šis metodas yra atsparesnis įvairiems įrašymo kanalų iškraipymams. Dėl to žadinimo signalo pagrindinis dažnis plačiai naudojamas teismo tyrimuose, kur skirtingi įrašymo kanalai yra viena iš pagrindinių problemų. Šiuo metu pagrindinio tono pasiskirstymas dažniausiai modeliuojamas naudojant histogramas bei skaičiuojant įvairius atstumus ar dviejų histogramų atitikimus. Tačiau pagrindinio tono pasiskirstymas nėra gausinis ir histogramų vaizdas bei palyginimo rezultatai priklauso nuo panaudoto klasių skaičiaus. Šiame darbe pagrindinio tono pasiskirstymui išreikšti naudojome Gauso mišinių modelius (GMM), o įrašams palyginti skaičiavome atstumus bei GMM aproksimacijų atitikimus. Geriausi rezultatai gauti naudojant Kullbacko ir Leiblerio atstumą. Il. 8, bibl. 12, lent. 2 (anglų kalba; santraukos anglų ir lietuvių k.).