New Feature Vector Extraction Method for Speaker Recognition

Lyudmila Sukhostat[1], Yadigar Imamverdiyev[2]
Institute of Information Technology of ANAS, Baku, Azerbaijan
[1]*lsuhostat@hotmail.com*, [2]*yadigar@lan.ab.az*

***Abstract*— Speech signal contains information not only connected to the pronounced phrase, but also data about speaker, language, environment, emotional state of the speaker. The main objective of the research is development of methods and algorithms increasing the precision of speaker recognition preserving acceptable indicators on computational complexity. Extraction of vectors of speech signal is an important stage of speaker recognition. Method based on Hilbert-Huang transform considering instability and non-linearity of human speech, as well as effective noise cancelling of the spectrum was proposed in the article.**

Keywords – speaker recognition, spectral features of speech signal, Hilbert-Huang transform

I. INTRODUCTION

Extraction of features is one of the important problems in speaker recognition systems. However, a priori it is impossible to evaluate which features fit better for recognition. Matching feature definition process consists of searching of possible option followed with experimental evaluation.

In connection with this, works on search for informative features of speech signal providing its adequate description and low percentage of errors during its recognition remain relevant.

Currently, there is no formal procedure of receiving a system of informative speech signal features providing qualitative speaker recognition. Usually they are chosen exclusively based on experience and intuition of the specialist. Then, more economic and informative subsystem of speech signal description is selected out of received features based on one or another formal method.

Research of physics of vocal apparatus [1], peripheral auditory system, experiments on reading dynamic spectrogram of speech signal called visible speech, and different psychophysical experiments demonstrate that information transfer in speech signal is realized by changes in its short-time amplitude spectrum reflecting articulation process.

Objective of feature allocation consists of transform of speech signal to a certain type of parametric representation for further analysis and processing. Speech signal slowly changes with time.

During the checking, within quite a short period of time (from 5 to 100 ms), its characteristics are sufficiently stationary. However, over a long period of time (about 0.2 seconds or more) characteristics of the signal change depending on the way speech sounds. Therefore, short-time spectral features are most frequently applied in tasks for speaker and speech recognition. Unlike high-level features requiring more complicated pre-processing [1, 2], they are easier to distinct and obtain good results [3].

Most popular approaches for speech signal feature extraction are methods based on Fourier, wavelet-analysis, as well as linear prediction method. Yet, they do not consider instability and non-linearity of human speech. In given article, method of speech signal feature extraction for speaker recognition tasks based on Hilbert transform is proposed.

II. SUMMARY OF SPEECH SIGNAL FEATURE EXTRACTION METHODS

In accordance with [4], majority of features can be classified: spectral features, spectral-time features, speech source features, high-level features and prosodic features. Among them, we can select Linear Prediction Coefficients (LPC) [5], Linear Prediction Cepstral Coefficients (LPCC) [6], Mel-Frequency Cepstral Coefficients (MFCC) which were applied to speaker recognition for the first time by Furui [7] and others.

MFCC features are most known and popular spectral features. MFCC and LPCC features were primarily developed for speech recognition and based on linear model source-filter for speech-generating system. During recent years, an interest was generated for non-linear models of speech-generation and works dedicated to features based on non-linear models (Teager energy operator based cepstral coefficients and amplitude-frequency modulation (AM-FM) based '*Q*' features) lead to significant improvement of features of speaker recognition [8].

These features match well when speech samples for education and recognition are clean (without noise) and recorded in similar conditions. Experiments show that information that is specific for the speaker is contained not only in frequencies traditionally used for speech and speaker recognition (lower than 4 kHz), but as well as higher frequencies (between 4 and 8 kHz). MFCC is capable of capturing information contained in high-frequency components. Bank of filters, used in MFCC allows low frequency components better than high-frequency.

A method based on Fourier, wavelet-analysis, as well as linear prediction method does not consider instability and non-linearity of speech signals. A method based on Hilbert-Huang

transform is proposed for negotiation of above-mentioned complications.

III. Feature Extraction Method Based on Hilbert-Huang Transform

Hilbert-Huang Transform (HHT) is regarded as Empirical Mode Decomposition (EMD) of non-linear and non-stable processes and Hilbert Spectral Analysis (HSA) [9]. HHT represents a time-and-frequency data analysis and does not require a prior functional basis transform. Instantaneous frequencies are calculated from derivatives of phase functions using Hilbert basis function transform.

EMD method is applied in virtue of non-stability and non-linearity of speech signals. Each of mode oscillations represents an Intrinsic Mode Functions (IMFs), defined by following rules:

1) Quantity of function extremes (maximums and minimums) and quantity of zeros must not differ for more than a unit.

2) At any point average values of envelopes built on local extremes are equal to zero.

EMD divides speech signal *x(t)* to IMFs set applying iterative Sifting Process.

EMD Process includes following stages:

Step 1: Identify all local extremes of the signal by coordinates and amplitudes. Accept $i = 1$.

Step 2: IMF Extraction procedure:

a) Calculate upper and lower envelopes of the signal using cubic (or some other) spline by extracted maximum and minimum. Determine the function of average values $m_1(t)$ between envelopes and find first approximation to first IMF mode functions:

$$h_1(t) = x(t) - m_1(t). \quad (1)$$

b) Repeat steps 1 and 2, considering functions $h_1(t)$ instead of $x(t)$ and find second approximation to IMF – first mode function $h_{11}(t)$

$$h_{11}(t) = h_1(t) - m_{11}(t). \quad (2)$$

c) Analogically, find the third and next approximations to IMF first mode function.

$$h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t). \quad (3)$$

d) Last value of $h_{1k}(t)$ iterations is accepted as high-frequency mode function $c_1(t) = h_{1k}(t)$ of IMFs family, which is directly included in original signal $x(t)$. This allows to calculate $c_1(t)$ from the content of the signal and leave lower frequency contents. We receive the residue:

$$r_1(t) = x(t) - c_1(t). \quad (4)$$

e) Accept $i = i + 1$ and go to step 1.

Step 3. In such manner, decomposition of signal in n-mode empirical approximation is achieved in the amount with residue $r_n(t)$:

$$X(t) = \sum_{j=1}^{n} c_j(t) + r_n(t) \quad (5)$$

Thus, next step of HHT is Hilbert transform. Using of this transform for each IMF allows to receive value of instantaneous frequency and amplitude for each moment of time. Let's describe application of Hilbert transform in more detail. Transform is applied with each IMF $c_j(t)$ in order to obtain $H[c_j(t)]$:

$$H[c_j(t)] = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{c_j(\tau)}{t - \tau} d\tau \quad (6)$$

And we can build analytical signal $Z_j(t)$ as

$$Z_j(t) = c_j(t) + i\, H[c_j(t)] = \alpha_j(t) \exp(i\theta_j(t)). \quad (7)$$

Amplitude function $\alpha_j(t)$ and phase function $\theta_j(t)$ changing in time are determined as follows:

$$\alpha_j(t) = \sqrt{c_j^2(t) + H^2[c_j(t)]}, \quad (8)$$

$$\theta_j(t) = \arctan \frac{H[c_j(t)]}{c_j(t)}. \quad (9)$$

Instantaneous value of frequency of non-stationary signal can be calculated as following:

$$\omega_j(t) = \frac{d\theta_j(t)}{dt} \quad (10)$$

Thus, after application of Hilbert transform to each IMF, original signal can be expressed as a real part of following expression:

$$X(t) = \mathrm{Re} \sum_{j=1}^{n} \alpha_j(t) \exp[i\theta_j(t)] = \mathrm{Re} \sum_{j=1}^{n} \alpha_j(t) \exp[i \int \omega_j(t) dt]. \quad (11)$$

Then the marginal spectrum and the marginal energy spectrum can be determined as:

$$h(\omega) = \int_0^T H(\omega, t)\, dt \quad (12)$$

$$E(\omega) = \int_0^T H^2(\omega, t)\, dt, \quad (13)$$

where T - is sampling length. Marginal spectrum expresses the measure of double amplitude or energy, obtained from each frequency.

HHT meets the requirements of adaptivity for analysis of non-stationary signals. Thus, signal can be locally and accurately reflected in temporary frequency domain by using Hilbert spectrum. On the other hand, EMD allows to dynamically extracting features of the signal depending on

oscillations present in the signal. EMD is more effective for noise cancelling that simple frequencies filter.

IV. Experimental Results

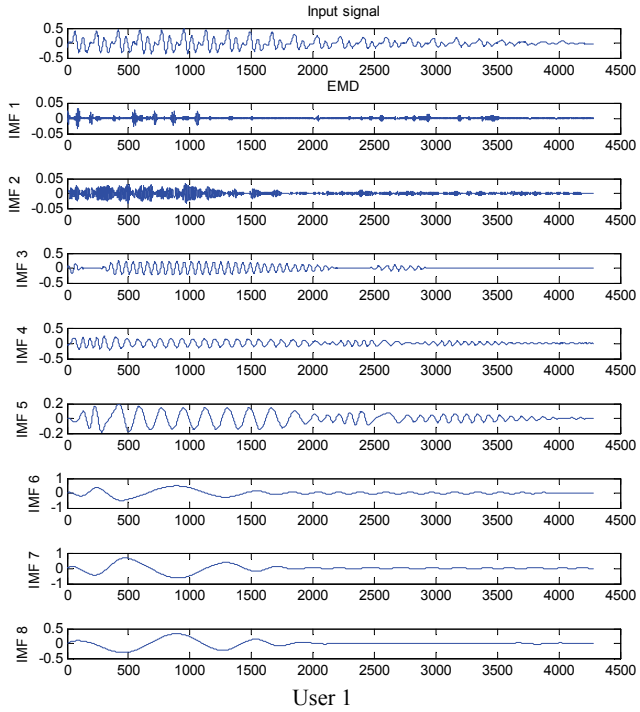
Unique characteristics are extracted from speech signal with considering that each speaker has individual characteristics. This means that IMF of one speaker differs from another.

Experiment was conducted on speech samples obtained from 20 speakers. Following combinations of samples were taken for experiment: 1) male and female, 2) female and female, and 3) male and male. Results of conducted experiments were provided on Fig. 1, which demonstrates incoming signal and received IMFs after application of EMD (using Azeri word "bir" (one) as a sample). We used first eight IMFs, as calculation process takes less time. It is seen from the figures, that IMFs of each signal differ from each other. Feature vector for each speaker is as following:

$C_{si} = [c_1(i), c_2(i), \ldots, c_8(i)]$, where $i = \overline{1,20}$.

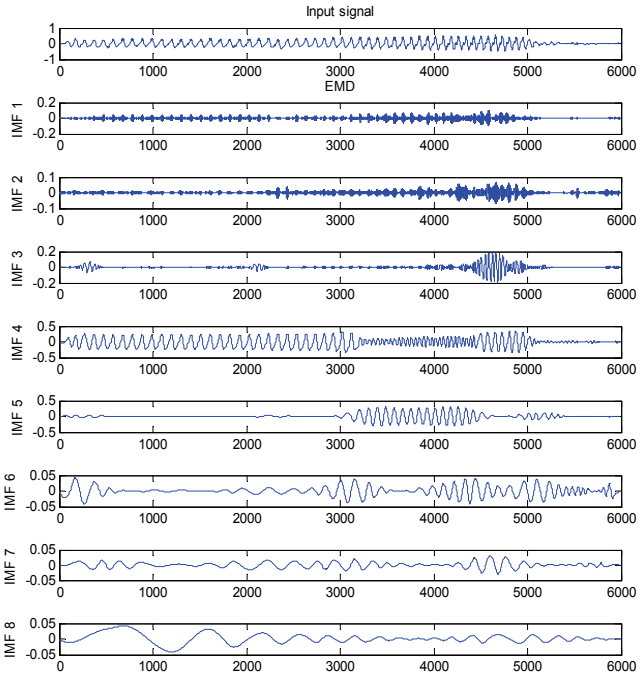
Spectrum received after HHT for word "bir" is given on Fig. 2.

Experiments were conducted on Matlab R2011b.



User 1

a)



User 2

b)

Figure 1. Decomposition of signals using EMD for word "bir". IMF – time and frequency components of a signal. IMF1 has the highest frequency spectrum.

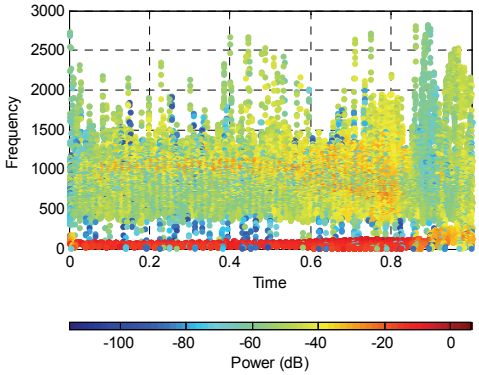


Figure 2. Spectrum received after Hilbert transformation.

V. Conclusion

It is known, the most known approaches to speech signal feature extraction based on Fourier, wavelet-analysis, and linear prediction method. However, they do not consider non-stability and non-linearity of human speech.

Approach to speaker recognition based on HHT is described in the article. Advantage of this method is that signal segmentation to words is not required, which significantly reduces processing time. The proposed method is the most accurate time-frequency representation of speech signal parameters in comparison with traditional spectral methods. It preservers internal characteristics of data and proposes a data presentation without limitations of uncertainty

principle. It is simple in realization and gives physically significant results in real time. In additional, it generates IMF through adaptive algorithm of a data set, which does not operate with other methods. Applying EMD, we receive IMF, which are the unique characteristics of each speaker.

ACKNOWLEDGMENT

This work was supported by the Science Development Foundation under the President of the Republic of Azerbaijan - Grant № EİF-2011-1(3)-82/08/1.

REFERENCES

[1] J. Benesty, M. Sondhi, Y. Huang, "Springer handbook of speech processing", Springer, 2007.

[2] G. Doddington, "Speaker recognition based on idiolectal differences between speakers", Proc. of Eurospeech, vol. 4, pp. 2521-2524, 2001.

[3] D. Reynolds, "Channel robust speaker verification via feature mapping", Proc. of ICASSP, vol. 2, pp. 53–56, 2003.

[4] T. Kinnunen, "Spectral features for automatic text-independent speaker recognition", Licentiate thesis, Department of Computer Science, University of Joensuu, Joensuu, Finland, 2003.

[5] J.D. Markel, A.H. Gray, "Linear Prediction of Speech", Springer, 1976.

[6] T. Kinnunen, H. Li, "An overview of text-independent speaker recognition: from features to supervectors", Speech Communication, vol. 52, no. 1, pp. 12-40, 2010.

[7] S. Furui, "Cepstral analysis techniques for automatic speaker verification", IEEE tran. acoust., speech, signal processing, vol. 27, pp. 254-272, 1981.

[8] A. Neustein, H. A. Patil, "Forensic Speaker Recognition", Springer, 2012.

[9] N.E. Huang, "Hilbert-Huang Transform and its applications", World Scientific Publishing, 2005.