Lyudmila V. Sukhostat DOI: 10.25045/jpit.v08.i1.06
Institute of Information Technology, National Academy of Sciences of Azerbaijan
lsuhostat@hotmail.com

ADAPTIVE NOISE REDUCTION METHOD BASED ON EMPIRICAL WAVELET TRANSFORM

Biometric user authentication by voice is one of the most important functions of the information security. But, changes in the acoustic environment and communication channels create noise and various distortions in speech signals, whereby the recognition accuracy in such systems is considerably degraded. Therefore, removal of noise in speech signals is essential to improve the accuracy of speaker recognition. This paper proposes a method of adaptive noise reduction based on empirical wavelet transform, which is tested on speech signals with different noise levels.

Keywords: *speech signal features, wavelets, empirical wavelet transform, discrete energy separation algorithm.*

Introduction

Channel distortion is a serious problem for the verification and identification of the speaker, as even an insignificant amount of distortion in a speech may change the unique features of the human voice, and destroy important information.

There are currently a large number of methods for speaker recognition based on several approaches, which show promising results for the identification and verification of a person's voice. However, the issue of establishing high accuracy systems for speaker identification still remains unsolved.

The main purpose of research in the field of speaker recognition is the development of methods and algorithms that improve the recognition accuracy, maintaining acceptable performance on computational complexity, at the same time.

At present, wavelet transform is widely used and applied in various signals processing to highlight specific features.

Wavelets are adapted to the signal, characterizing its local properties. Unlike other known methods of working with signaling, such as windowed Fourier transform, wavelets have several advantages:

- availability of windows of variable size for flexible analysis;
- providing time-frequency information about a signal;
- application of band-pass filter bank.

Thus, the approach of synchrosqueezing, proposed in [1, 2], is based on the selection of appropriate wavelets. It removes insignificant wavelet coefficients (in time and scale) taking into account the threshold value of corresponding to this portion of the signal.

To achieve the same goals, other recent works include Empirical Wavelet Transform (EWT) for the construction of adaptive wavelet basis for the expansion of a given signal to adaptive "bands". [3] This model is drawn upon a robust pre-treatment for detecting peaks, and then, performs a spectral segmentation based on the detected peaks. It generates a corresponding wavelet filter bank. The filter bank is flexible at a spectral overlapping.

This adaptive method leads to signal degradation at its basic mode. The method combines the strength of wavelet formalism and adaptability of the Empirical Mode Decomposition (EMD). Indeed, current degradation models are mainly limited to:

- algorithmic nature, which lacks mathematical theory (EMD);
- recursive screenings of most techniques, which do not allow the reverse error correction;
- fail to properly deal with the noise;
- strict limits of wavelet approaches.

EWT is one of adaptive methods used where the basis is dependent on the information content of the signal. The method has a great advantage in the extraction of steady and unsteady components from the signal. Here, the mode is presented as the components AM-FM [4]. The segmentation of the Fourier spectrum is performed, and a small filtration is used for the detection of the reference limits. Orthonormal basis is formed on the basis of the information provided in the signal.

Given all abovementioned, this paper proposes a new approach for the extraction of speech signal features based on EWT.

A new approach to extracting speech signal features based on empirical wavelet transform

Accordingly, EWT and an inverse transform are applied for signal analysis. EWT enables to separate both approximating and detailing factors [5]. It is based on a priori filtering and has a strong mathematical basis. EWT is widely applied to non-stationary signals for noise reduction.

First, the Fourier spectrum of the speech signal is determined by the interval $[0,\pi]$, and then, it is divided into different segments N. The boundaries between the segments are denoted by ω_n, $n=\overline{0,N}$ (where $\omega_0=0$ and $\omega_N=\pi$). A high or the highest signal peak is found. The mid-point is considered in the middle of the two peaks. A filter bank is created. The output filter is convoluted with the original speech signal to produce the Intrinsic Mode Functions (IMF). Initially, noisy IMFs are obtained, which are subjected to filtration to clean speech signal from the noise.

The segmentation is carried out to produce "hard" frames, in order to avoid data loss. The following conditions should be provided for this:

$$\gamma < \min_n\left(\frac{\omega_{n+1}-\omega_n}{\omega_{n+1}+\omega_n}\right).$$

The empirical scalable function and empirical wavelets are considered and defined as follows in the case study:

$$\hat{\phi}_n(\omega)=\begin{cases}1, & if \quad |\omega|\le(1-\gamma)\omega_n\\ \cos\left[\frac{\pi}{2}\beta\left(\frac{2}{\gamma\omega_n}\left(|\omega|-(1-\gamma)\omega_n\right)\right)\right], & if \quad (1-\gamma)\omega_n<|\omega|\le(1+\gamma)\omega_n\\ 0, & otherwise\end{cases} \tag{1}$$

and

$$\hat{\psi}_n(\omega)=\begin{cases}1, & if \quad (1+\gamma)\omega_n\le|\omega|\le(1-\gamma)\omega_{n+1}\\ \cos\left[\frac{\pi}{2}\beta\left(\frac{2}{\gamma\omega_n}\left(|\omega|-(1-\gamma)\omega_{n+1}\right)\right)\right], & if \quad (1-\gamma)\omega_{n+1}<|\omega|\le(1+\gamma)\omega_{n+1}\\ \sin\left[\frac{\pi}{2}\beta\left(\frac{2}{\gamma\omega_n}\left(|\omega|-(1-\gamma)\omega_n\right)\right)\right], & if \quad (1-\gamma)\omega_n\le|\omega|\le(1+\gamma)\omega_n\\ 0, & otherwise\end{cases} \tag{2}$$

where $0<\gamma<1$ for all $n>0$.

Arbitrary function $\beta(x)$ of $C^k([0,1])$, that

$$\beta(x)=\begin{cases}3x^2-2x^3, & 0\le x\le 1\\ 0, & x<0\\ 1, & 1<x\end{cases} \tag{3}$$

if $\beta(x)=1-\beta(1-x)$ for all $x\in[0,1]$.

With the decreasing filter width, i.e. with the decreasing scaling parameter its amplitude is also increased.

We can now determine empirical wavelet transform $W_f^{\varepsilon}(n,t)$ as in case of classical wavelet transform:

$$W_f^{\varepsilon}(n,t) = \langle f, \psi_n \rangle = \int f(\tau)\overline{\psi_n(\tau - t)}d\tau = \hat{f}(\omega) * \overline{\hat{\psi}_n(\omega)} \ , \quad (4)$$

and present approximating coefficients as scalar products with scalable function

$$W_f^{\varepsilon}(0,t) = \langle f, \phi_1 \rangle = \int f(\tau)\overline{\phi_1(\tau - t)}d\tau = \hat{f}(\omega) * \overline{\hat{\phi}_1(\omega)} \ , \quad (5)$$

where $\hat{\phi}_1(\omega)$ and $\hat{\psi}_n(\omega)$ are determined by the equations (1) and (2) respectively. Inverse transform is as follows

$$f(t) = W_f^{\varepsilon}(0,t) * \phi_1(t) + \sum_{n=1}^{N} W_f^{\varepsilon}(n,t) * \psi_n(t) = \hat{W}_f^{\varepsilon}(0,\omega) * \hat{\phi}_1(\omega) + \sum_{n=1}^{N} \hat{W}_f^{\varepsilon}(n,\omega) * \hat{\psi}_n(\omega) \ . \quad (6)$$

IMF function f_k is defined as follows:

$$f_0(t) = W_f^{\varepsilon}(0,t) * \phi_1(t) \quad (7)$$

$$f_k(t) = W_f^{\varepsilon}(k,t) * \psi_k(t) \, . \quad (8)$$

Finding the IMF, Discrete Energy Separation Algorithm (DESA) is used [6], which has a low computational complexity, and quickly runs on real signals. The method decompose the speech signal into AM and FM components (Fig. 1).

$d^m(n)$ denotes the IMF value for each frame when $n = 1, \ldots, N$ and $m = 1, \ldots, M_x$, where M_x is the number of modes, $x(t)$ is split into which.

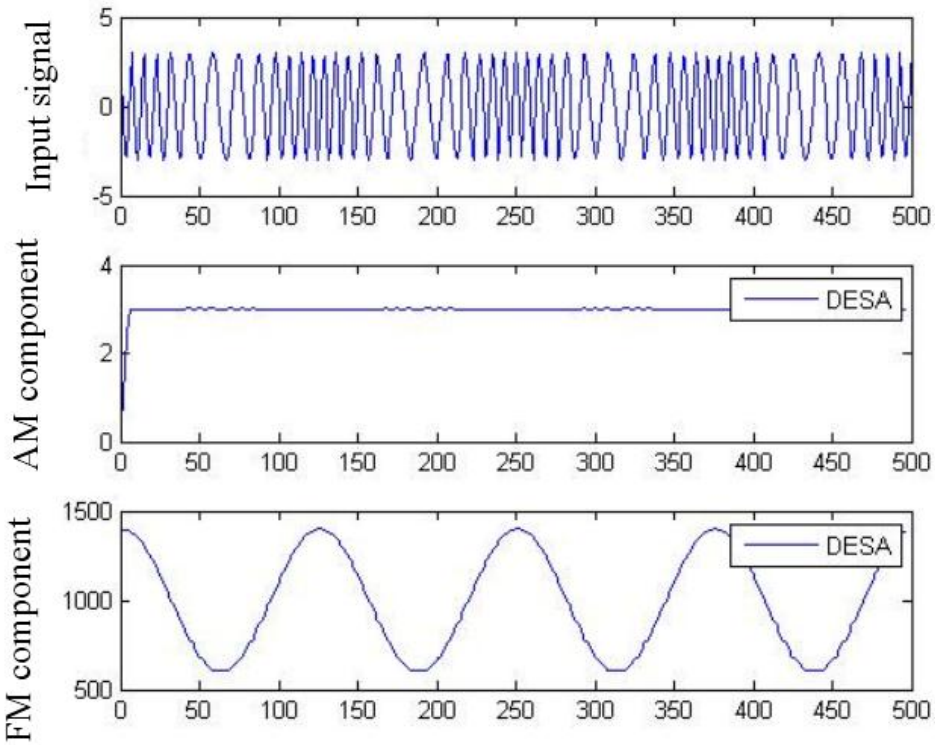

Fig.1. Obtaining AM and FM components with DESA

Then discrete operator Teager can be applied:

$$\Psi\left[d^m(n)\right]=\left(d^m(n)\right)^2-d^m(n-1)d^m(n+1),\quad n=\overline{2,N-1}, \tag{9}$$

if $d^m(n)$ discrete cosine with constant amplitude A and frequency ω, $d^m(n)=A\cos(\Omega n+\theta)$ when $\Omega=\omega T$ and T denotes discrete period,

$$\Psi\left[d^m(n)\right]=A^2\omega^2\left(\frac{\sin\Omega}{\Omega}\right)^2. \tag{10}$$

The instantaneous frequency $\Omega(n)$ and instantaneous amplitude $a(n)$ are considered as the parameters:

$$\Omega(n)=\arccos\left(1-\frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[d^m(n)]}\right), \tag{11}$$

$$|a(n)|=\sqrt{\frac{\Psi\left[d^m(n)\right]}{1-\left(1-\frac{\Psi\left[y(n)\right]+\Psi\left[y(n+1)\right]}{4\Psi\left[d^m(n)\right]}\right)^2}}, \tag{12}$$

where $y(n)=d^m(n)-d^m(n-1)$ for $n=2,\ldots,N$.

The instantaneous frequency and amplitude are combined to produce a short-term instantaneous frequency estimation of mean amplitude F_{mean} for each speech signal [7] under

$$F_{mean}=\frac{\sum_{n=2}^{N}\Omega(n)a(n)}{\sum_{n=2}^{N}a(n)}. \tag{13}$$

The equation (11) provides more accurate estimation, and more robustness to low energy and noisy frequency bands.

Experimental results

MATLAB system is used for the experimental studies, which includes additional Wavelet Toolbox for the analysis of speech signals. The speech database for the Azerbaijani language is chosen for the experiments [8]. SNR is calculated using the developed method based on the modified EWT at various percentages of noise signals. Productivity of the proposed approach is compared to the EWT method (Table 1).

Table 1

Performance Comparison of feature extraction methods

SNR (dB) / Method	-5	0	10	15
EWT-DESA	15,11	5,77	10,56	4,79
Proposed approach	14,79	5,12	9,98	3,96

The method showed decreasing percent of Gross Pitch Error (GPE) with the increasing noise level. Thus, the resulting scalable function is highly sensitive to short-term high-frequency signal changes, thereby, reduces the noise level.

Conclusion

Whilst processing the voice signals, the noise removal is one of the important problems.

The only way to overcome this is the models that are constantly adapting to the speech signal changes.

Taking into account all the above mentioned, the proposed method based on EWT shows quite promising results at different noise levels, which enables to identify a number of areas for further research: selection of the basic functions and transformation types to improve the accuracy of the proposed method, and its application in the field of voice recognition.

References

1. Daubechies I., Lu J., Wu H.-T. Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool // Applied and Computational Harmonic Analysis, 2010, vol.30, no.2, pp.243–261.
2. Wu H.-T., Flandrin P., Daubechies I. One or Two Frequencies? The Synchrosqueezing Answers // Advances in Adaptive Data Analysis, 2011, vol.3, no.1–2, pp.29–39.
3. Gilles J. Empirical Wavelet Transform // IEEE Transactions on Signal Processing, 2013, vol.61, no.16, pp.3999–4010.
4. Holambe R.S., Deshpande M.S. Noise robust speaker identification: using nonlinear modeling // Forensic Speaker Recognition, 2012, pp.153–182
5. Imamverdiev Y.N., Suhostat L.V. Razrabotka robastnogo metoda izvlecheniya rechevyh priznakov na osnove ehmpiricheskogo vejvlet preobrazovaniya // Informacionnye tekhnologii, 2015, №1, c. 19–23.
6. Schlotthauer G., Torres M.E., Rufiner H.L. A new algorithm for instantaneous F0 speech extraction based on Ensemble Empirical Mode Decomposition / Proc. of 17th European Signal Processing Conf., 2009, pp.2347–2351.
7. Chhabra S., Bajaj R., Pachori R.B., Biswas R.N. Features based on Fourier-Bessel expansion for application of speaker identification system / Proc. Of Indian Conf. for Academic Research by Undergraduate Students, 2010, pp.1–3.
8. Imamverdiyev Y.N., Sukhostat L.V. AZ-SRDAT - speech database for the Azerbaijan language // Problems of Information Technology, 2013, No 1, pp. 67-73.