УДК 004.934

Я. Н. Имамвердиев[1], канд. техн. наук, науч. сотр., зав. отд., e-mail: yadigar@lan.ab.az,
Л. В. Сухостат[2], e-mail: lsuhostat@hotmail.com,
Институт информационных технологий НАНА, Баку, Азербайджан

Разработка робастного метода извлечения речевых признаков на основе эмпирического вейвлет-преобразования

Извлечение векторов признаков речевого сигнала является важным этапом для систем распознавания диктора. В настоящее время остаются актуальными работы по поиску информативных признаков речевого сигнала, обеспечивающих его адекватное описание и низкий процент ошибок при распознавании. В данной работе представлен подход для извлечения речевых признаков на основе эмпирического вейвлет-преобразования, повышающего точность распознавания, сохраняя при этом приемлемые показатели по вычислительной трудоемкости.

Ключевые слова: *распознавание диктора, эмпирическое вейвлет-преобразование, дискретный алгоритм разделения энергии, мгновенная амплитуда, мгновенная частота*

Введение

Существует множество голосовых признаков, характеризующих диктора. Речь является сложным сигналом, возникающим в результате нескольких преобразований, происходящих на различных уровнях: семантическом, лингвистическом, артикуляционном и акустическом [1]. Различия, связанные с диктором, являются результатом сочетания анатомических различий, присущих вокальному тракту и манере разговора разных людей. Все эти различия могут быть использованы при распознавании диктора.

В настоящее время не существует формальной процедуры получения системы информативных признаков речевого сигнала, обеспечивающих качественное распознавание диктора. Обычно их выбирают исключительно на основе опыта и интуиции специалиста. Затем из полученной таким образом исходной системы признаков тем или иным формальным способом выбирают более экономичную и наиболее информативную подсистему описания речевого сигнала.

Исследования физики голосового аппарата [2], периферической слуховой системы, опытов по чтению динамических спектрограмм речевого сигнала, называемых видимой речью, и различных психофизических экспериментов показывают, что передача информации в речевом сигнале реализуется изменениями его кратковременного амплитудного спектра.

Цель выделения признаков заключается в преобразовании сигнала речи к некоторому типу параметрического представления для дальнейшего анализа и обработки. Кратковременные спектральные признаки наиболее часто применяют в задачах распознавания диктора и речи. В отличие от признаков высокого уровня, требующих более сложной предварительной обработки [2, 3], их легче вычислить и получить хорошие результаты [4]. Кепстральные признаки тесно связаны с лингвистическим содержанием речи. Помимо кепстральных особенностей, речь имеет и источник возбуждения, который, как полагают, содержит полезные свойства для распознавания диктора. Кроме того, в реальных ситуациях существуют большие различия между этапами разработки и практического применения системы распознавания диктора. Как следствие, кепстральные признаки недостаточны, чтобы обеспечить удовлетворительную и надежную точность распознавания диктора. Они также не учитывают нестационарность и нелинейность человеческой речи.

Данная работа направлена на исследование новых и эффективных параметров для робастного распознавания диктора и предлагает метод извлечения признаков речевого сигнала для задачи распознавания диктора на основе эмпирического вейвлет-преобразования (Empirical Wavelet Transform, EWT).

1. Краткий обзор методов извлечения признаков речевого сигнала

Многие исследования были посвящены разработке различных схем извлечения характерных для диктора акустических признаков из речевых высказываний. J. Wolf в работе [5] среди наиболее существенных параметров выделяет частоту основного тона, спектральные признаки гласных и назальных согласных, оценку голосового источника, продолжительность слова и время начала "озвончения" (*voice onset time*). В работе [6] авторы сделали обзор и обобщили основные особенности речи, которые были использованы для системы распознавания диктора. Наряду с классическими и ведущими признаками были приведены некоторые недавно полученные наборы параметров.

Свойства идеальных речевых признаков, применяемых в системах распознавания диктора [5, 7]:

- большая вариабельность между дикторами и небольшая изменчивость у каждого диктора;
- устойчивость к фоновому шуму и искажениям;
- частое использование в обычной речи;
- простота в измерении;
- стабильность во времени и независимость от здоровья/настроения говорящего;
- трудно имитируемые.

Признаки, как правило, классифицируются на пять групп с точки зрения их физической интерпретации [8]: спектральные, спектрально-временные признаки голосового источника, просодические и признаки высокого уровня.

Имеется множество спектральных признаков, характеризующих речевой сигнал в задачах распознавания речи и диктора. Среди них можно выделить коэффициенты линейного предсказания (*linear prediction coefficients*, LPC) [9], кепстральные коэффициенты линейного предсказания (*linear prediction cepstral coefficients*, LPCC), кепстральные коэффициенты по шкале мел (*mel-frequency cepstral coefficients*, MFCC), которые были впервые применены к распознаванию диктора [10] и другие.

MFCC-признаки являются наиболее известными и популярными спектральными признаками. MFCC- и LPCC-признаки первоначально были разработаны для распознавания речи и основаны на линейной модели источник — фильтр для системы речеобразования.

В последние годы при описании и анализе свойств речи был использован подход на основе AM — FM (Amplitude Modulation — Frequency Modulation) моделирования.

Монокомпонент AM — FM сигнала описывается уравнением

$$x(n) = A(n)\cos[\Theta_n], \qquad (1)$$

где A(n) — мгновенная амплитуда монокомпонентного сигнала; Θ_n — мгновенная фаза. При этом многокомпонентный сигнал сначала разлагается, каждый его компонент описывается мгновенной огибающей и мгновенной частотой. Данный подход показывает значительные улучшения показателей распознавания диктора [11].

2. Эмпирическое вейвлет-преобразование

Мы предлагаем метод построения семейства вейвлетов, адаптированных к обрабатываемым сигналам. Один из путей в достижении адаптивности состоит в предположении, что фильтры зависят от расположения информации в спектре анализируемого сигнала. Для четкости рассмотрим реальные сигналы (где спектр симметричен относительно частоты $\omega = 0$), но подобные рассуждения могут

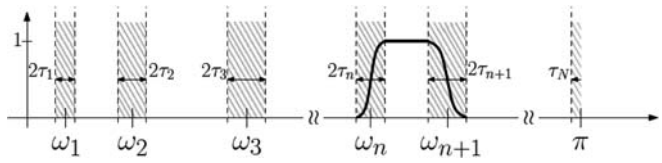

Рис. 1. Разбивка оси Фурье

быть также расширены для комплексных сигналов путем построения различных фильтров на положительных и отрицательных частотах. Мы также рассмотрим нормализованную ось Фурье, которая имеет периодичность для того, чтобы удовлетворить критерию Шеннона, и ограничим наше обсуждение отрезком $\omega \in [0, \pi]$.

Начнем с предположения, что отрезок $[0, \pi]$ делится на N смежных сегментов. Обозначим границы между сегментами ω_n (где $\omega_0 = 0$ и $\omega_N = \pi$), как показано на рис. 1. Каждый сегмент $\Lambda_n = [\omega_{n-1}, \omega_n]$. Вокруг ω_n определяем переходную фазу T_n шириной $2\tau_n$.

Эмпирические вейвлеты [12] определяются как полосовые фильтры на каждом Λ_n. Для этого используем идею, применяемую при построении вейвлетов Littlewood-Paley и Meyer. Тогда для всех $n > 0$ определяем эмпирическую масштабируемую функцию $\hat{\phi}_n$ и эмпирические вейвлеты согласно следующим выражениям:

$$\hat{\phi}_n(\omega) = \begin{cases} 1, \text{ если } |\omega| \leqslant \omega_n - \tau_n; \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}(|\omega| - \omega_n + \tau_n)\right)\right], \\ \text{если } \omega_n - \tau_n \leqslant |\omega| \leqslant \omega_n + \tau_n; \\ 0, \text{ в противном случае,} \end{cases} \qquad (2)$$

$$\hat{\psi}_n(\omega) = \begin{cases} 1, \text{ если } \omega_n + \tau_n \leqslant |\omega| \leqslant \omega_{n+1} - \tau_{n+1}; \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_{n+1}}(|\omega| - \omega_{n+1} + \tau_{n+1})\right)\right]; \\ \text{если } \omega_{n+1} - \tau_{n+1} \leqslant |\omega| \leqslant \omega_{n+1} + \tau_{n+1} \\ \sin\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}(|\omega| - \omega_n + \tau_n)\right)\right], \\ \text{если } \omega_n - \tau_n \leqslant |\omega| \leqslant \omega_n + \tau_n; \\ 0, \text{ в противном случае.} \end{cases} \qquad (3)$$

Функция $\beta(x)$ из C^k ([0, 1]) (пространство k раз дифференцируемых функций на интервале [0, 1] удовлетворяет условию

$$\beta(x) = \begin{cases} 0, \text{ если } x \leqslant 0; \\ 1, \text{ если } x \geqslant 1. \end{cases}$$
$$\text{и } \beta(x) + \beta(1 - x) = 1 \ \forall x \in [0, 1]. \qquad (4)$$

Что касается выбора τ_n, возможны несколько вариантов. Самый простой состоит в выборе τ_n пропорционально ω_n: $\tau_n = \gamma\omega_n$, где $0 < \gamma < 1$. Следова-

тельно, для всех $n > 0$ уравнения (2) и (3) принимают вид

$$\hat{\phi}_n(\omega) = \begin{cases} 1, \text{ если } |\omega| \leqslant (1-\gamma)\omega_n; \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}(|\omega| - (1-\gamma)\omega_n)\right)\right], \\ \text{если } (1-\gamma)\omega_n \leqslant |\omega| \leqslant (1+\gamma)\omega_n; \\ 0, \text{ в противном случае} \end{cases} \quad (5)$$

и

$$\hat{\psi}_n(\omega) = \begin{cases} 1, \text{ если } (1+\gamma)\omega_n \leqslant |\omega| \leqslant (1-\gamma)\omega_{n+1}; \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_{n+1}}(|\omega| - (1-\gamma)\omega_{n+1})\right)\right], \\ \text{если } (1-\gamma)\omega_{n+1} \leqslant |\omega| \leqslant (1+\gamma)\omega_{n+1}; \\ \sin\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}(|\omega| - (1-\gamma)\omega_n)\right)\right], \\ \text{если } (1-\gamma)\omega_n \leqslant |\omega| \leqslant (1+\gamma)\omega_n; \\ 0, \text{ в противном случае.} \end{cases} \quad (6)$$

Теперь можем определить эмпирическое вейвлет-преобразование $W_f^{\varepsilon}(n, t)$ так же как и в случае классического вейвлет-преобразования.

Результаты представлены в виде скалярных произведений с эмпирическими вейвлетами

$$W_f^{\varepsilon}(n, t) = \langle f, \psi_n \rangle = \int f(\tau)\overline{\psi_n(\tau - 1)}\, d\tau = \\ = \hat{f}(\omega)\overline{\psi_n(\omega)}, \quad (7)$$

а аппроксимирующие коэффициенты — в виде скалярных произведений с масштабируемой функцией

$$W_f^{\varepsilon}(0, t) = \langle f, \phi_1 \rangle = \int f(\tau) \cdot \overline{\phi_1(\tau - t)}\, d\tau = \\ = \hat{f}(\omega) \cdot \overline{\hat{\phi}_1(\omega)}, \quad (8)$$

где $\hat{\psi}_n(\omega)$ и $\hat{\phi}_1(\omega)$ определяются из уравнений (5) и (6) соответственно. Обратное преобразование принимает вид

$$f(t) = W_f^{\varepsilon}(0, t) * \phi_1(t) + \sum_{n=1}^{N} W_f^{\varepsilon}(n, t) * \psi_n(t) = \\ = \hat{W}_f^{\varepsilon}(0, \omega) * \hat{\phi}_1(\omega) + \sum_{n=1}^{N} \hat{W}_f^{\varepsilon}(n, \omega) * \hat{\psi}_n(\omega). \quad (9)$$

Эмпирическая мода (*Intrinsic Mode Function*, IMF) f_k определяется следующим образом:

$$f_0(t) = W_f^{\varepsilon}(0, t) * \phi_1(t), \quad (10)$$

$$f_k(t) = W_f^{\varepsilon}(k, t) * \psi_k(t). \quad (11)$$

Примеры получения IMF-компонент с помощью EWT показаны на рис. 2 и 3. Первый тестовый сигнал $f_{sig1}(t)$ (рис. 2) получен путем суммирования трех компонент (для $t \in [0, 1]$)

$$f_{s1}(t) = 6t^2; \quad (12)$$

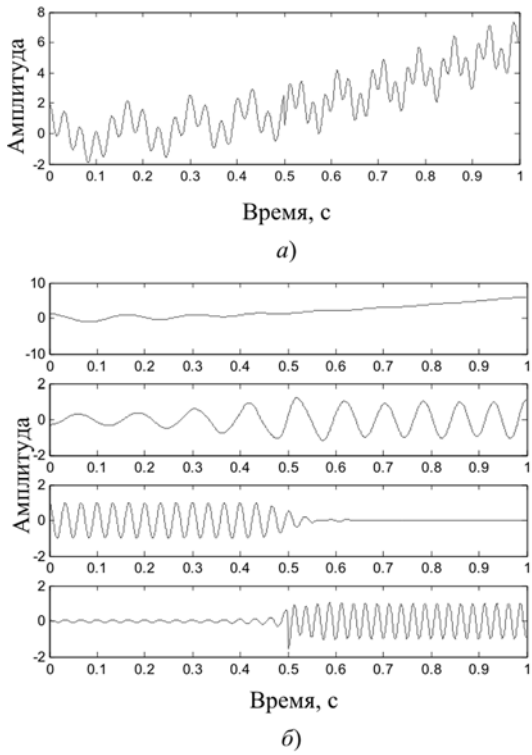

Рис. 2. Входной сигнал $f_{sig1}(t)$ (*а*) и эмпирические моды, полученные путем EWT (*б*)

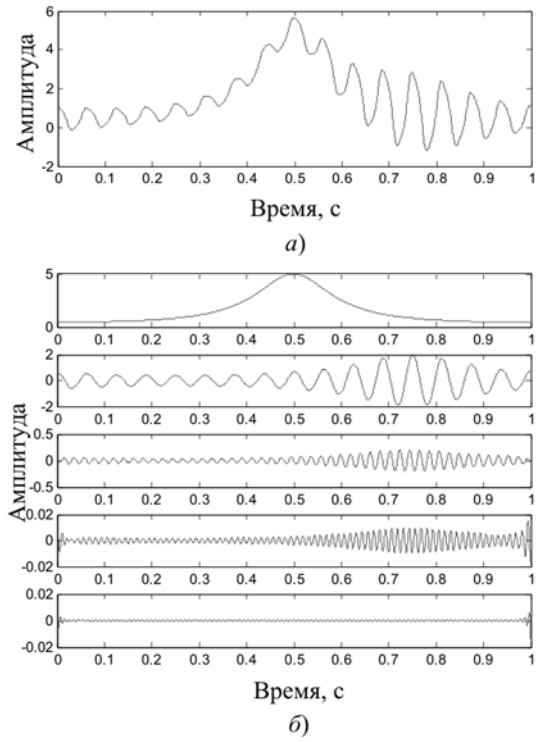

Рис. 3. Входной сигнал $f_{sig2}(t)$ (*а*) и эмпирические моды, полученные путем EWT (*б*)

$$f_{s2}(t) = \cos(10\pi t + 10\pi t^2); \quad (13)$$

$$f_{s3}(t) = \begin{cases} \cos(80\pi t - 15\pi), \text{ если } t > 0{,}5; \\ \cos(60\pi t), \text{ в противном случае.} \end{cases} \quad (14)$$

$$f_{sig1}(t) = f_{s1}(t) + f_{s2}(t) + f_{s3}(t). \quad (15)$$

Компоненты второго сигнала $f_{sig2}(t)$ принимают вид

$$f_{s1}(t) = \frac{1}{1{,}2 + \cos(2\pi t)}; \quad (16)$$

$$f_{s2}(t) = \frac{1}{1{,}5 + \sin(2\pi t)}; \quad (17)$$

$$f_{s3}(t) = \cos(32\pi t + \cos(64\pi t)) \quad (18)$$

и

$$f_{sig2}(t) = f_{s1}(t) + f_{s2}(t) f_{s3}(t). \quad (19)$$

3. Дискретный алгоритм разделения энергии

После получения IMF необходимо выбрать метод для выделения мгновенной амплитуды и частоты. Обычно применяется преобразование Гильберта [13], однако дискретный алгоритм разделения энергии (*Discrete Energy Separation Algorithm*, DESA) [14] превосходит его по вычислительной сложности и быстроте на реальных сигналах. А также рассматривает энергию, необходимую для генерации каждого монокомпонента AM — FM сигнала. Это помогает при исследовании вибраций голосовых складок для выделения отличительных особенностей каждого диктора.

Пусть $d^m(n)$ — значение IMF для каждого фрейма при $n = 1, ..., N$ и $m = 1, ..., M_x$, где M_x обозначает число мод, на которые $x(t)$ разбивается.

Затем мы можем применить дискретный оператор Тигера

$$\Psi[d^m(n)] = (d^m(n))^2 - d^m(n-1)d^m(n+1), \quad n = 2, ..., N-1. \quad (20)$$

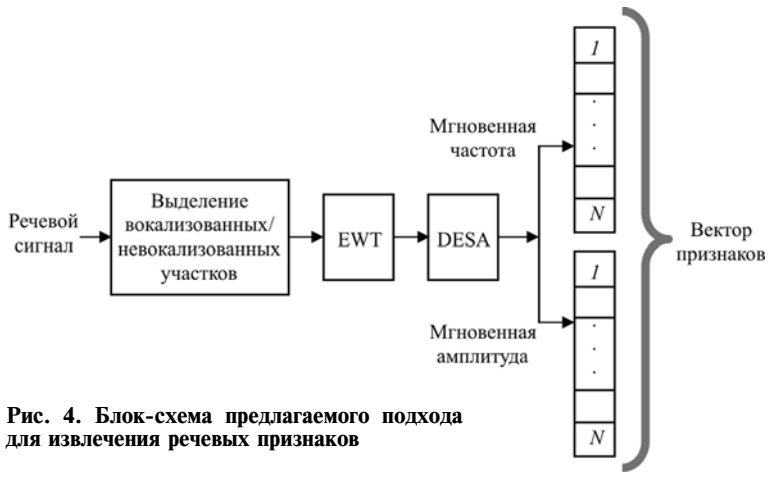

Рис. 4. Блок-схема предлагаемого подхода для извлечения речевых признаков

Если $d^m(n)$ — дискретный косинусный с постоянной амплитудой A и частотой ω, $d^m(n) = A\cos(\Omega n + \theta)$ при $\Omega = \omega T$ и T — период дискретизации, то

$$\Psi[d^m(n)] = A^2\omega^2\left(\frac{\sin\Omega}{\Omega}\right)^2. \quad (21)$$

Затем применим алгоритм DESA для AM — FM разделения. Он оценивает мгновенную частоту $\Omega(n)$ и мгновенную огибающую $a(n)$ следующим образом:

$$\Omega(n) = \arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[d^m(n)]}\right), \quad (22)$$

$$|a(n)| = \sqrt{\frac{\Psi[d^m(n)]}{1 - \left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[d^m(n)]}\right)^2}}, \quad (23)$$

где $y(n) = d^m(n) - d^m(n-1)$ для $n = 2, ..., N$.

4. Извлечение речевых признаков на основе предлагаемого подхода

На рис. 4 показана блок-схема получения признаков, характерных для каждого диктора, путем модуляции источника возбуждения. Приводятся компоненты мгновенной частоты и мгновенной амплитуды, а также вектор, объединяющий все эти компоненты.

Процесс вычисления вектора признаков следующий:

1. *Выделение вокализованных/невокализованных участков*. Вектор признаков извлекается только из вокализованных участков.

2. *Получение IMF-компонент на основе EWT:*

a. Применение преобразования Фурье.

b. Вычисление локального максимума на отрезке $[0, \pi]$ и нахождение множества $\{\omega_n\}$.

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

d. Построение банка фильтров.

e. Фильтрация сигнала для получения каждого компонента IMF.

3. *Применение DESA*. Нахождение мгновенной частоты и мгновенной амплитуды на основе алгоритма выделения энергии Тигера.

4. *Объединение векторов признаков*. Полученные вектора мгновенных частот и мгновенных амплитуд далее объединяются для получения нового вектора признаков.

Заключение

В работе описан подход для распознавания диктора, основанный на эмпирическом вейвлет-преобразовании. Он дает физически

значимые результаты в режиме реального времени. Генерирует IMF через адаптивный алгоритм из набора данных. Применяя EWT, мы получаем IMF, которые являются уникальными особенностями для каждого диктора. Для выделения мгновенной частоты и мгновенной амплитуды был применен алгоритм DESA. Он превосходит преобразование Гильберта, сохраняя внутренние свойства данных, не ограничиваясь принципом неопределенности.

Данная работа выполнена при финансовой поддержке Фонда развития науки при Президенте Азербайджанской Республики — грант № EİF-RİTN-MQM-2/İKT-2-2013-7(13)-29/18/1.

Список литературы

1. **Quatieri T. F.** Discrete-time speech signal processing: principles and practice, ser. NJ: Prentice Hall, 2001. 819 p.
2. **Benesty J., Sondhi M., Huang Y.** Springer handbook of speech processing. Springer, 2008. 1159 p.
3. **Doddington G.** Speaker recognition based on idiolectal differences between speakers // Proc. of Eurospeech. 2001. V. 4. P. 2521—2524.
4. **Reynolds D.** Channel robust speaker verification via feature mapping // Proc. of ICASSP. 2003. V. 2. P. 53—56.
5. **Wolf J. J.** Efficient acoustic parameters for speaker recognition // J. Acoustical Society of America. 1982. V. 51, N. 6 (Part 2). P. 2044—2056.
6. **Kinnunen T., Li H.** An overview of text-independent speaker recognition: from features to supervectors // Speech Communication. 2010. V. 52, N. 1. P. 12—40.
7. **Rose P.** Forensic speaker identification / Taylor & Francis forensic science series. New York: Taylor & Francis, 2002. 380 p.
8. **Kinnunen T.** Spectral features for automatic text-independent speaker recognition: Licentiate thesis. Department of Computer Science. University of Joensuu. Finland, 2003.
9. **Маркел Дж., Грей А. Х.** Линейное предсказание речи. М.: Связь, 1980. 308 с.
10. **Furui S.** Cepstral analysis techniques for automatic speaker verification // IEEE tran. acoust., speech, signal processing. 1981. V. 27. P. 254—272.
11. **Holambe R. S., Deshpande M. S.** Noise Robust Speaker Identification: Using Nonlinear Modeling // Forensic Speaker Recognition. 2012. P. 153—182.
12. **Gilles J.** Empirical Wavelet Transform // IEEE Transactions on Signal Processing. 2013. V. 61. N. 16. P. 3999—4010.
13. **Huang N. E.** Hilbert-Huang Transform and its applications. World Scientific Publishing, 2005. 311 p.
14. **Schlotthauer G., Torres M. E., Rufiner H. L.** A new algorithm for instantaneous F0 speech extraction based on Ensemble Empirical Mode Decomposition // 17th European Signal Processing Conf. 2009. P. 2347—2351.

Y. N. Imamverdiyev, Head of Department, e-mail: yadigar@lan.ab.az,
L. V. Sukhostat, Researcher, e-mail: lsuhostat@hotmail.com,
Institute for Information Technologies, Azerbaijan National Academy of Sciences, Baku, Azerbaijan

Development of Robust Speech Feature Extraction Method Based on Empirical Wavelet Transform

Speech feature vectors extraction is an important step for speaker recognition systems. Currently, state-of-art works remain relevant to find informative features of speech signals, ensuring its appropriate description and low error rate during recognition. In this paper we present an approach for speech feature extraction based on empirical wavelet transform. To calculate the instantaneous frequency and instantaneous amplitude of IMFs Discrete Energy Separation Algorithm is used, which overcomes the disadvantages of Hilbert transform. The proposed method increases the recognition accuracy, while maintaining an acceptable level of computational complexity.

Keywords: *speaker recognition, empirical wavelet transform, discrete energy separation algorithm, instantaneous amplitude, instantaneous frequency*

References

1. **Quatieri T. F.** *Discrete-time speech signal processing: principles and practice, ser.* NJ: Prentice Hall, 2001. 819 p.
2. **Benesty J., Sondhi M., Huang Y.** *Springer handbook of speech processing.* Springer, 2008. 1159 p.
3. **Doddington G.** Speaker recognition based on idiolectal differences between speakers // *Proc. of Eurospeech.* 2001. V. 4. P. 2521—2524.
4. **Reynolds D.** Channel robust speaker verification via feature mapping // *Proc. of ICASSP.* 2003. V. 2. P. 53—56.
5. **Wolf J. J.** Efficient acoustic parameters for speaker recognition // *J. Acoustical Society of America.* 1982. V. 51. N. 6 (Part 2). P. 2044—2056.
6. **Kinnunen T., Li H.** An overview of text-independent speaker recognition: from features to supervectors // *Speech Communication.* 2010. V. 52, N. 1. P. 12—40.
7. **Rose P.** *Forensic speaker identification.* Taylor & Francis forensic science series. New York: Taylor & Francis. 2002. 380 p.
8. **Kinnunen T.** *Spectral features for automatic text-independent speaker recognition.* Licentiate thesis. Department of Computer Science. University of Joensuu. Finland. 2003.
9. **Markel Dzh., Grej A. X.** *Linejnoe predskazanie rechi.* M.: Svjaz', 1980. 308 p.
10. **Furui S.** Cepstral analysis techniques for automatic speaker verification // *IEEE tran. acoust., speech, signal processing.* 1981. V. 27. P. 254—272.
11. **Holambe R. S., Deshpande M. S.** Noise Robust Speaker Identification: Using Nonlinear Modeling // *Forensic Speaker Recognition.* 2012. P. 153—182.
12. **Gilles J.** Empirical Wavelet Transform // *IEEE Transactions on Signal Processing.* 2013. V. 61, N. 16. P. 3999—4010.
13. **Huang N. E.** *Hilbert-Huang Transform and its applications.* World Scientific Publishing, 2005. 311 p.
14. **Schlotthauer G., Torres M. E., Rufmer H. L.** A new algorithm for instantaneous F0 speech extraction based on Ensemble Empirical Mode Decomposition // *Proc. 17th European Siignal Processing Conf.* 2009. P. 2347—2351.