УДК 004.934.8'1

Имамвердиев Я.Н.[1], Сухостат Л.В.[2]
Институт Информационных Технологий НАНА, Баку, Азербайджан
[1]yadigar@lan.ab.az, [2]lsuhostat@hotmail.com

ОБ ОДНОМ МЕТОДЕ ИЗВЛЕЧЕНИЯ ПРИЗНАКОВ ДЛЯ СИСТЕМ РАСПОЗНАВАНИЯ ДИКТОРА

Создание методов, повышающих точность работы систем распознавания диктора, является основной целью исследований. Извлечение векторов признаков является важным этапом для систем распознавания диктора. Для извлечения признаков в работе предлагается метод на основе преобразования Гильберта-Хуанга, учитывающий нестационарность и нелинейность человеческой речи.

Ключевые слова: распознавание диктора, извлечение признаков речевого сигнала, преобразование Гильберта-Хуанга, внутренняя модовая функция.

Введение

Процесс определения подходящих признаков заключается в переборе возможных вариантов с последующей экспериментальной оценкой.

В связи с этим остаются актуальными работы по поиску информативных признаков речевого сигнала, обеспечивающих низкий процент ошибок при распознавании.

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 выбирается более экономичная и наиболее информативная подсистема описания речевого сигнала.

Исследования физики голосового аппарата [1] показали, что анатомические характеристики дикторов для определения производительности систем распознавания с женскими голосами значительно хуже систем с мужскими голосами [2, 3].

Цель выделения признаков заключается в преобразовании сигнала речи к некоторому типу параметрического представления для дальнейшего анализа и обработки. Речевой сигнал медленно меняется со временем. При проверке в течение достаточно короткого периода времени (от 5 до 100 мс) его характеристики достаточно стационарны. Тем не менее, в течение длительного периода времени (порядка 0,2 сек и более) характеристики сигнала изменяются в зависимости от того, как звучит речь. Таким образом, кратковременные спектральные признаки наиболее часто применяются в задачах распознавания диктора и речи. В отличие от признаков высокого уровня, требующих более сложной предварительной обработки [1], их легче вычислить и получить хорошие результаты [4].

Большинство признаков речевого сигнала не учитывают нестационарность и нелинейность человеческой речи. В данной работе предлагается метод извлечения признаков речевого сигнала для задачи распознавания диктора на основе преобразования Гильберта-Хуанга (Hilbert-Huang Transform, HHT).

Метод извлечения признаков на основе преобразования Гильберта-Хуанга

Имеется множество спектральных признаков, характеризующих речевой сигнал в задачах распознавания речи и диктора. Среди них можно выделить коэффициенты линейного предсказания (Linear Prediction Coefficients, LPC) [5], кепстральные коэффициенты линейного предсказания (Linear Prediction Cepstral Coefficients, LPCC) [6],

кепстральные коэффициенты по шкале мел (Mel-Frequency Cepstral Coefficients, MFCC) [7], которые были впервые применены к распознаванию диктора Furui [8], и другие.

Однако данные методы не учитывают нестационарность и нелинейность речевых сигналов. Рассмотрим метод, основанный на преобразовании Гильберта-Хуанга.

Под HHT понимается метод эмпирической модовой декомпозиции (Empirical Mode Decomposition, EMD) нелинейных и нестационарных процессов и Гильбертов спектральный анализ (Hilbert Spectral Analysis, HSA) [9]. HHT представляет собой частотно-временной анализ данных и не требует априорного функционального базиса преобразования (Рис. 1). Мгновенные частоты (Instantaneous Frequencies, IFs) вычисляются от производных фазовых функций Гильбертовым преобразованием функций базиса.

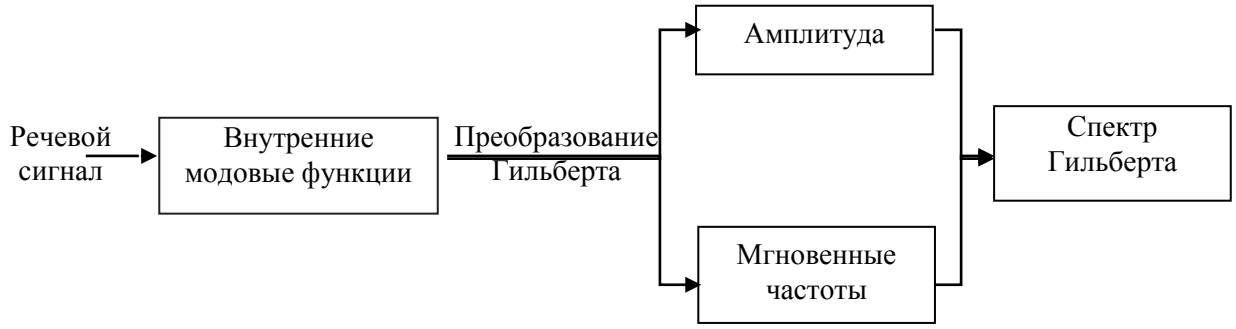

Рис.1. Общая схема HHT

Метод EMD [10] применяется в силу нестационарности и нелинейности речевых сигналов.

Схема процесса EMD представлена на рис. 2. Процедура позволяет вычислить внутренние модовые функции (Intrinsic Mode Functions, IMFs). IMFs – частотно-временная компонента сигнала. Первая IMF имеет наиболее высокий частотный спектр

Следующим шагом преобразования Гильберта-Хуанга является преобразование Гильберта. Использование преобразования для каждой IMF позволяет получить значения мгновенной частоты и амплитуды для каждого момента времени. Опишем подробнее применение преобразования Гильберта. Оно применяется к каждой IMF $c_j(t)$, чтобы получить $H[c_j(t)]$:

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

и мы можем построить аналитический сигнал $Z_j(t)$ как

$$Z_j(t) = c_j(t) + i\,H[c_j(t)] = \alpha_j(t)\exp(i\theta_j(t)). \tag{2}$$

Так определяются изменяющаяся во времени функция амплитуды $\alpha_j(t)$ и фазовая функция $\theta_j(t)$:

$$\alpha_j(t) = \sqrt{c_j^2(t) + H^2[c_j(t)]} \tag{3}$$

$$\theta_j(t) = \arctan\frac{H[c_j(t)]}{c_j(t)} \quad . \tag{4}$$

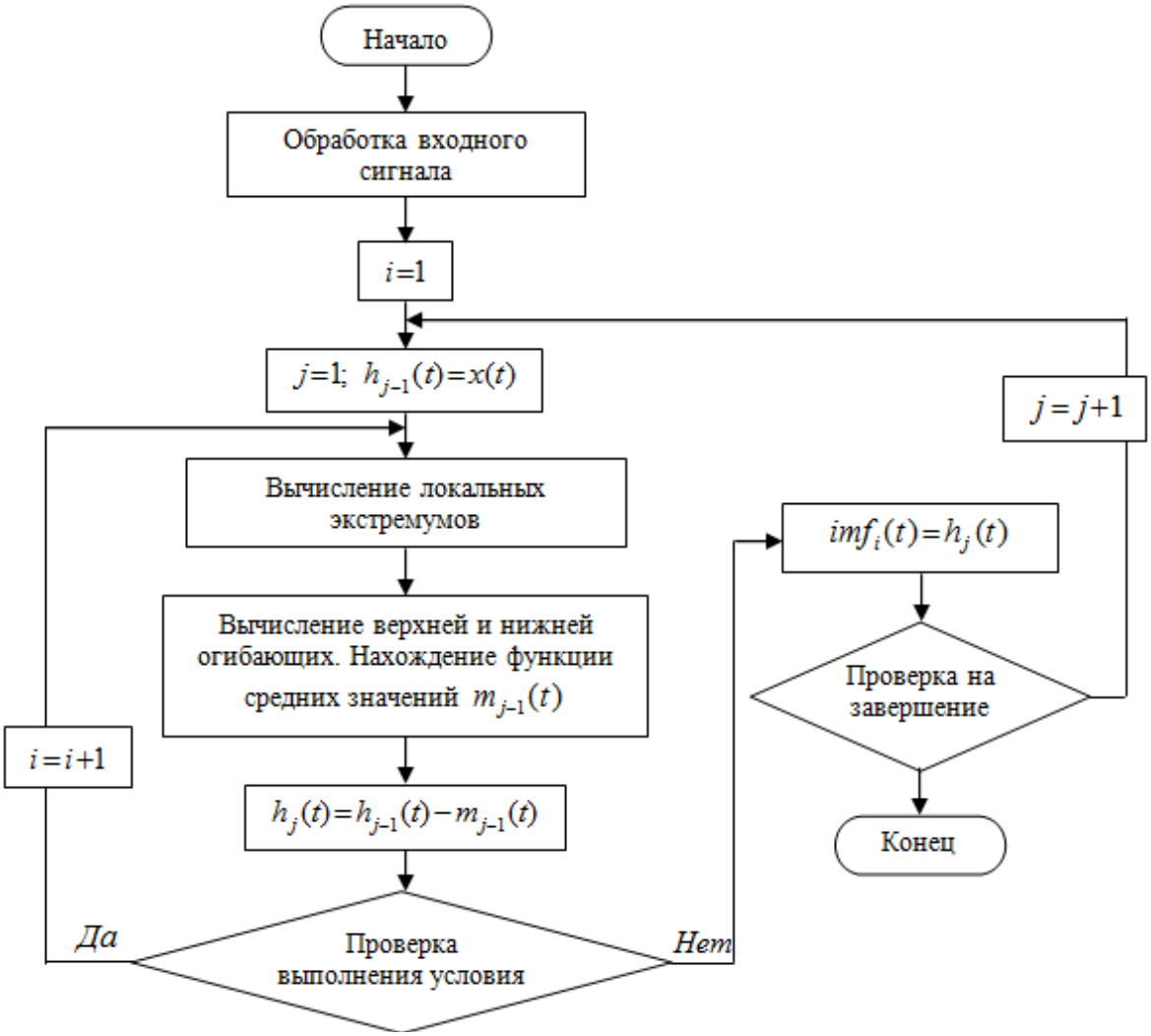

Рис. 2. Процесс эмпирической модовой декомпозиции

Мгновенное значение частоты нестационарного сигнала может быть вычислено следующим образом:

$$\omega_j(t)=\frac{d\theta_j(t)}{dt}. \tag{5}$$

Таким образом, после применения преобразования Гильберта к каждой IMF исходный сигнал может быть выражен как реальная часть следующего выражения:

$$X(t)=\mathrm{Re}\sum_{j=1}^{n}\alpha_j(t)\exp[i\theta_j(t)]=\mathrm{Re}\sum_{j=1}^{n}\alpha_j(t)\exp[i\int\omega_j(t)dt]. \tag{6}$$

Уравнение (6) выражено через $H(\omega,t)$, который представляет амплитуду и мгновенную частоту как функцию времени в трехмерном графике. Тогда предельный (пограничный) спектр может быть определен как

$$h(\omega)=\int_0^T H(\omega,t)\,dt \tag{7}$$

$$E(\omega) = \int_0^T H^2(\omega, t)\, dt, \qquad (8)$$

где T – длина сигнала (sampling length). Предельный (пограничный) спектр выражает меру удвоенной амплитуды или энергию, получаемую из каждой частоты.

EMD позволяет динамически извлекать признаки сигнала в зависимости от помех, присутствующих в сигнале. Он является более эффективным, чем обычный частотный фильтр для подавления шумов.

Результаты экспериментов

Уникальные характеристики извлекаются из речевого сигнала с учетом того, что каждый диктор имеет свои индивидуальные особенности. Что означает, что IMF одного диктора отличается от другого.

Эксперимент проводился на речевых образцах из базы данных TIMIT [11]. Для эксперимента были взяты 20 образцов мужских и женских голосов. Поскольку вычислительный процесс занимает большое количество времени, были рассмотрены первые восемь IMF. Из рис. 3 (на примере слова bir) видно, что IMF для мужского и женского голосов отличаются друг от друга. Вектор признаков для каждого диктора представляет собой вектор $C_{si} = [c_1(i), c_2(i), \ldots, c_8(i)]$, где $i = \overline{1,20}$.

Эксперименты проводились на Matlab R2011b.

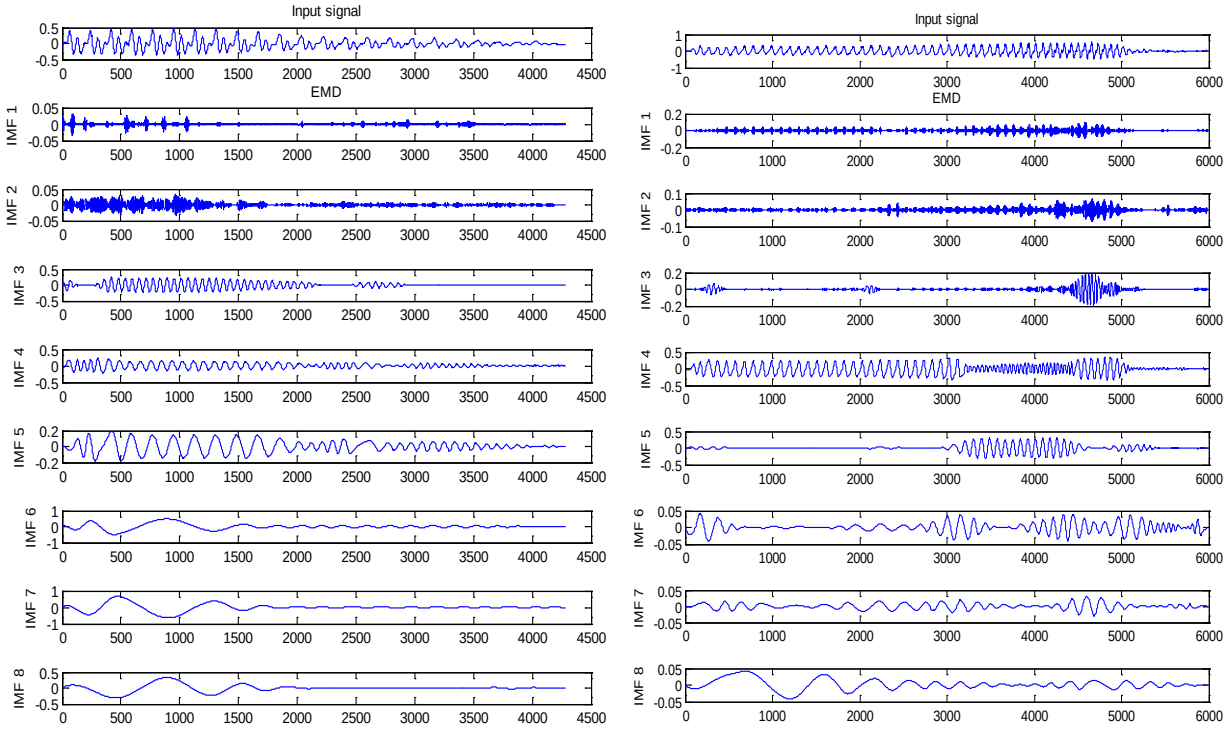

Рис.3. Декомпозиция сигналов с помощью EMD
для мужского и женского голосов на примере слова «bir»

В таблице 1 приводится сравнение полученных вероятности пропуска «своего» (Genuine Acceptance Rate, GAR) и вероятности пропуска «чужого» (False Acceptance Rate, FAR) для мужских и женских голосов с данными, полученными на конкурсе NIST Speaker Recognition Evaluation, проводимом Национальным Институтом Стандартов и Технологий (National Institute of Standards and Technology, NIST) [12] начиная с 1996 г.

Результаты показывают более высокую точность распознавания для мужских голосов. Объяснение этому состоит в том, что в расслабленном состоянии речевой тракт лучше измеряется у мужчин, т.к. период колебания голосовых связок у них длиннее, чем у женщин.

Таблица 1

Результаты проведенных экспериментов

	GAR	FAR
Мужские голоса	94,75%	18,05%
Женские голоса	94,67%	17,76%
Среднее значение	94,71%	17,91%
NIST SRE 2010	~85-90%	~10-15%

Заключение

В работе предложен метод извлечения признаков речевого сигнала для задачи распознавания диктора, основанный на преобразовании Гильберта-Хуанга. Преимуществом этого метода является то, что нет необходимости в сегментации сигнала на слова, тем самым значительно уменьшается время обработки. Данный метод является более точным частотно-временным представлением параметров речевого сигнала по сравнению с традиционными спектральными методами. Он сохраняет внутренние свойства данных, не ограничиваясь принципом неопределенности. Метод прост в реализации и дает физически значимые результаты в режиме реального времени. Генерирует IMF через адаптивный алгоритм из набора данных, с которым другие методы не работают. Применяя EMD, получаем IMF, которые являются уникальными особенностями для каждого диктора.

Литература

1. Benesty J., Sondhi M., Huang Y. Springer handbook of speech processing. Springer, 2007, 1176 p.
2. Johnson K., Speaker Normalization in speech perception, In Pisoni, D.B. & Remez, R. (eds) The Handbook of Speech Perception. Oxford: Blackwell Publishers, 2005, pp. 363–389.
3. Mason J.S., Thompson J. Gender effects in speaker recognition // Proc. ICSP-93, Beijing, pp. 733–736, 1993.
4. Doddington G. Speaker recognition based on idiolectal differences between speakers // Proc. of Eurospeech, vol. 4, pp. 2521–2524, 2001.
5. Маркел Дж., Грей А.Х. Линейное предсказание речи. М.: Связь, 1980, 308 с.
6. Furui S. Cepstral analysis techniques for automatic speaker verification // IEEE tran. acoust., speech, signal processing, vol. 27, pp. 254–272, 1981.
7. Kinnunen T. Spectral features for automatic text-independent speaker recognition, Licentiate thesis, Department of Computer Science, University of Joensuu, Joensuu, Finland, 2003.
8. Kinnunen T., Li H. An overview of text-independent speaker recognition: from features to supervectors // Speech Communication, 2010, vol. 52, no. 1, pp. 12–40.
9. Huang N.E. Hilbert-Huang Transform and its applications, World Scientific Publishing, 2005.
10. Li X., Li D., Liang Z., Voss L.J., Sleigh J.W. Analysis of depth of anesthesia with Hilbert-Huang spectral entropy // Clin Neurophysiol, 2008, no. 119, pp. 2465–2475.
11. Ward N.C., Dersch D.R. Text-independent speaker identification and verification using the TIMIT database // Proc. of ICSLP, 1998, v. 2, pp. 233–237.

12. Przybocki M.A., Martin A.F. NIST Speaker Recognition Evaluation chronicles // Proc. Odyssey 2004: The Speaker and Language Recognition Workshop, Toledo, Spain, June 2004.

UOT 004.934.8'1

İmamverdiyev Yadigar N.[1], Suxostat Lyudmila V.[2]
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]yadigar@lan.ab.az. [2] lsuhostat@hotmail.com

Səsə görə şəxsin tanınması sistemləri üçün əlamətlərin çıxarılması metodu

Tanıma sistemlərinin dəqiqliyini artıran metodların yaradılması tədqiqatların əsas məqsədidir. Əlamət vektorlarının çıxarılması səsə görə şəxsin tanınması sistemlərində əhəmiyyətli mərhələdir. Məqalədə səsə görə şəxsin tanınması üçün qeyri-stasionar və insan nitqinin qeyri-xəttiliyini nəzərə alan Hilbert-Huanq çevirməsi əsasında metod təklif edilir.

Açar sözlər: *səsə görə şəxsin tanınması, nitq siqnalı əlamətlərinin çıxarılması, Hilbert-Huanq çevirməsi, daxili moda funksiyası.*

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

Feature vector extraction method for speaker recognition systems

Development of methods increasing the accuracy of speaker recognition system is the main objective of the research. Feature vector extraction is one of the most important stages of speaker recognition system. In this paper we propose a feature extraction method based on Hilbert-Huang transform considering instability and non-linearity of human speech.

Key words: *speaker recognition, speech feature extraction, Hilbert-Huang transform, Intrinsic Mode Function.*