УДК 004.934.8'1

СИСТЕМА ИДЕНТИФКАЦИИ ДИКТОРА В АКУСТИЧЕСКИХ ШУМАХ НА ОСНОВЕ АНТРОПОМОРФИЧЕСКОЙ ОБРАБОТКИ РЕЧЕВОГО СИГНАЛА

Д.Н. КРУЧОК, А.А. ПЕТРОВСКИЙ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 21 января 2016

Рассматривается система идентификации диктора в акустических шумах с использованием антропоморфической обработки речевого сигнала. Подробно описывается преобразование речевого сигнала на основе кохлеарной модели и его применение в задаче распознавания диктора. Полученный характеристический вектор на базе данного преобразования используется в качестве признаков для системы идентификации диктора. В качестве решающих правил применяются нейронные сети прямого распространения. Приводятся результаты распознавания разработанной системы идентификации диктора.

Ключевые слова: характеристический вектор, кохлеарная модель, система идентификации диктора.

Введение

Речевые технологии находят все большее распространение в различных сферах жизни человека: охранные системы и системы безопасности, криминалистика и судебная экспертиза, робототехника, системы управления оборудованием, средствах телекоммуникаций, криминалистика, банковский сектор и др. [1]. Во всех этих сферах, используемые компьютерные средства должны не только определить личность человека, вступающего в контакт с ними, но и уметь подстроится под него и под условия, в которых происходит общение. Согласно решению Federal Financial Institution Examination Council, USA, от 2005 года, использование однофакторной методологии аутентификации личности является неадекватным средством защиты в системах удаленного доступа к финансам [2]. Поэтому введение голосовой биометрии является целесообразным. Несмотря на то, что эффективность работы систем распознавания в закрытых помещениях с благоприятной акустической обстановкой находится на довольно высоком уровне, результаты распознавания данных систем в местах с наличием акустических шумов еще далеки от допустимых, и в своей работе требуют вмешательства человека. Таким образом, настоящее исследование ставит целью анализ результатов распознавания системы идентификации диктора в акустических шумах с использованием преобразования речевого сигнала на основе кохлеарной модели.

Системы идентификации диктора

Распознавание речи или диктора делится на текстозависимое (ограничение на то, что говорят), так и на текстонезависимое (нет ограничений на то, что говорят). В общем случае задача распознавания диктора подразделяется на два основных направления: идентификацию и верификацию [1]. При верификации пользователь предъявляет в том или ином виде свой идентификатор, и система распознавания должна подтвердить или отвергнуть этот идентификатор. При идентификации диктор не указывает своего идентификатора, и система распознавания должна установить, принадлежит ли речевой сигнал голосу одного из дикторов, которые имеются в базе.

В процессе идентификации диктора существует две отдельные стадии: обучение (или регистрация) и тестирование [3]. Во время обучения подготавливается речевая база, состоящая из выражений каждого известного диктора, который должен быть идентифицирован, для построения (обучения) модели для этого диктора. Обычно это стадия выполняется перед тем, как будет эксплуатироваться система идентификации.

Во время тестирования, неизвестное речевое выражение сравнивается с каждой натренированной моделью диктора. В идентификации с закрытым набором неизвестные личности принадлежат базе дикторов, которые участвовали в процессе обучения, и тогда проблема со случайным диктором, не участвующим в обучении отпадает. Главным показателем эффективности таких систем является точность распознавания (усредненный процент правильного распознавания среди всех дикторов в базе). Идентификация с закрытым набором обычно характерна для ведомственных организаций, где группы лиц известны, их данные могут быть собраны и храниться в базе данных, и для которых идентификация будет происходить внутри их ведомства или отдела (т.е. нет «внешних» пользователей).

В идентификации с открытым набором, неизвестным диктором может быть любой человек. Так как идентификация всегда осуществляется среди конечного, известного круга лиц, то невозможно идентифицировать случайного диктора. Поэтому первой задачей идентификации с открытым набором является определение, принадлежит ли диктор к базе известных дикторов, и если нет, то отклонить такого диктора, в противном случае – идентификация выполняется как с закрытым набором. Для таких систем важно определить принадлежность диктора к базе, иначе случайный диктор будет всегда идентифицироваться.

Общая структурная схема системы идентификации диктора представлена на рис. 1.

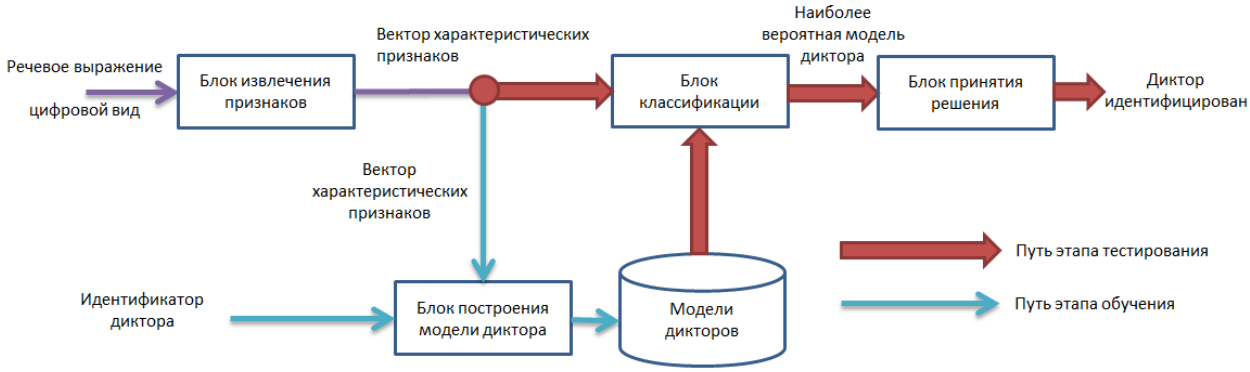


Рис. 1. Общая схема системы идентификации

Основными структурными блоками системы идентификации, которые отражают процесс обработки речевого сигнала, являются: блок экстракции характеристического вектора; блок построения (обучения) моделей дикторов; блок классификации (решающих правил); блок принятия решений. Первоначально, из речевого выражения, представленного в цифровом виде, извлекаются характеристические признаки, в пространстве которых будет происходить идентификация, содержащие ключевые особенности речи диктора. В процессе обучения, полученные признаки используются для построения модели известного диктора, а в процессе тестирования используются для классификации полученного речевого выражения. И затем принимается решение о результате идентификации. Блок экстракции характеристического вектора может быть реализован методом или алгоритмом экстракции признаков речевого сигнала: анализ и коэффициенты линейного предсказания (англ. LPC – Linear Prediction Coding) [3], анализ и коэффициенты линейного перцептуального предсказания (англ. PLP – Perceptual Linear Prediction) [4] и др. Блок построения (обучения) моделей и блок классификации связаны между собой и определяются выбранными решающими правилами, основными из которых являются: методы аппроксимации плотности вероятности в пространстве признаков взвешенной смесью нормальных распределений (англ. GMM – Gaussian Mixture Models), метод опорных векторов (англ. SVM – Support Vector Machines) и др. Блок принятия решения обычно определяет принадлежность речевого выражения к одной из моделей диктора, выбранной алгоритмом, если происходит идентификация с открытым набором. В противном случае блок выдает идентификатор (например, номер распознанного диктора).

Системы распознавания диктора, применяемые на практике, часто подвергаются зашумлению или искажению входных речевых выражений (ограничение по полосе пропускания в телефонном канале, наличие фоновых шумов среды и др.), что снижает их производительность. Решения этой проблемы, обеспечивающие шумоустойчивость (робастность) в идентификации диктора, делятся на следующие варианты: оценка вектора характеристических признаков на основе антропоморфической модели и выбор решающего правила [3].

В первом варианте шум убирается напрямую из характеристического вектора диктора. Этот метод включает в себя такие методы, как нормализация кепстрального среднего (англ. CMN – Cepstral Mean Normalization) [5], фильтрация модуляционного спектра (англ. RASTA – Relative Spectra Analysis) [6] и др. А методы, основанные на выборе решающего правила, пытаются включить искажающие характеристики и признаки шума в само решающее правило и тем самым достигнуть робастности системы идентификации. Таким образом, для построения системы идентификации диктора нужно решить две основные подзадачи: выбор характеристического вектора и выбор решающих правил.

Анализ акустических признаков. Преобразование речевого сигнала на основе кохлеарной модели

Индивидуальные особенности диктора зависят от строения речевого тракта, механики колебания голосовых складок и системы управления артикуляцией. Акустические признаки обязаны передавать особенности диктора, быть устойчивыми к различным искажениям, а также обладать компактностью представления для их использования в реальном масштабе времени. Для получения характеристических признаков оригинальный сигнал подвергается различным преобразованиям (преобразование Фурье, вейвлет-преобразование). Одним из таких преобразований также является преобразование на основе кохлеарной модели, предложенное в [7]. Данное преобразование моделирует импульсную характеристику базилярной мембраны в улитке человеческого уха и распределение ее нелинейных характеристик. В практических приложениях обработки цифровых сигналов преобразование применяется в дискретной форме [7]:

$$T[a_i, b] = \sum_{n=0}^{N} f[n] \frac{1}{\sqrt{|a_i|}} \psi\left[\frac{n-b}{a_i}\right], \qquad (1)$$

где $a_i = f_L / f_{C_i}$ – масштабирующий коэффициент для *i*-ой частотной полосы с центральной частотой f_{C_i}, может быть как в линейной, так и в нелинейной шкале частот; *N* – количество отсчетов цифрового сигнала $f[n]$.

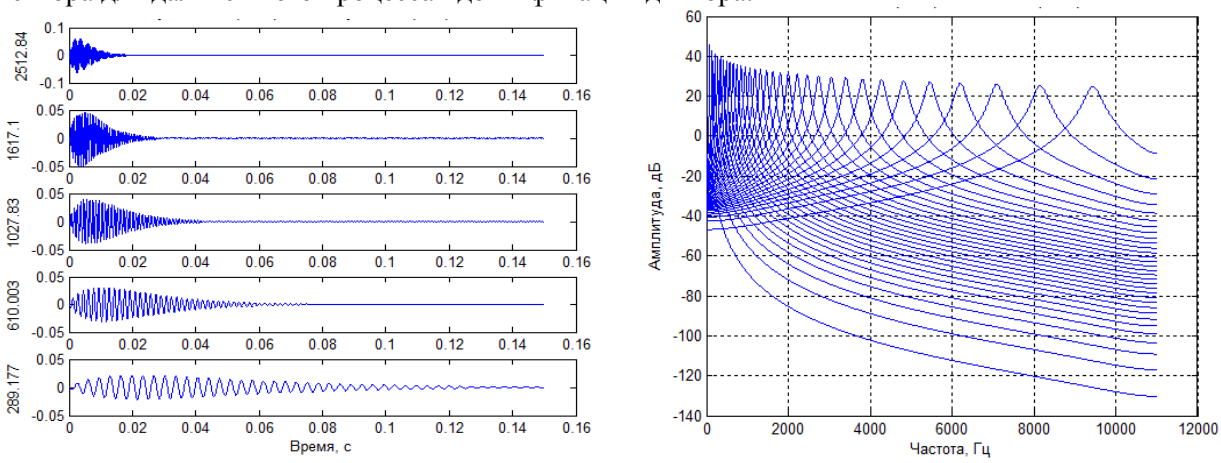
$$\psi\left[\frac{n-b}{a_i}\right] = \left[\frac{n-b}{a_i}\right]^{\alpha} \cdot \exp\left[-2\pi f_L \beta\left[\frac{n-b}{a_i}\right]\right] \cdot \cos\left[2\pi f_L\left[\frac{n-b}{a_i}\right] + \theta\right] \cdot u[n], \qquad (2)$$

где $\alpha > 0, \beta > 0, u[n]$ – функция Хевисайда; *b* – коэффициент сдвига.

Изменяя параметры α, β можно получить различные форму и ширину для каждого фильтра в полученном банке, которые будут наиболее точно соответствовать модели слуховой системы человека. Импульсные характеристики базилярной мембраны на разных частотах показаны на рис. 2, *а*, при этом $\alpha = 0{,}8; \beta = 0{,}02$. Указанные значения были выбраны как наиболее оптимальные для анализа речевого сигнала для дальнейшей идентификации диктора [8]. Метки по оси ординат слева от каждого графика соответствуют центральным частотам пяти фильтров (значения даны в герцах).

Такие результаты очень похожи на результаты физиологических экспериментов [7] и сопоставимы с кохлеарной моделью в [8], а это, в свою очередь, свидетельствует о том, что данное преобразование позволяет получить характеристический вектор признаков, который будет соответствовать характеристикам, которые выделяет слуховая система человека. На рис. 2, *б* представлены частотные характеристики используемого банка фильтров при значениях коэффициента $\beta = 0{,}02$ и $\alpha = 0{,}8$. В качестве тестового сигнала был взят речевой сигнал с частотой дискретизации 22050 Гц; количество используемых фильтров в банке равно 32. Так как

данное преобразование моделирует импульсную характеристику базилярной мембраны человеческого уха, то его целесообразно использовать для получения характеристического вектора для дальнейшего процесса идентификации диктора.

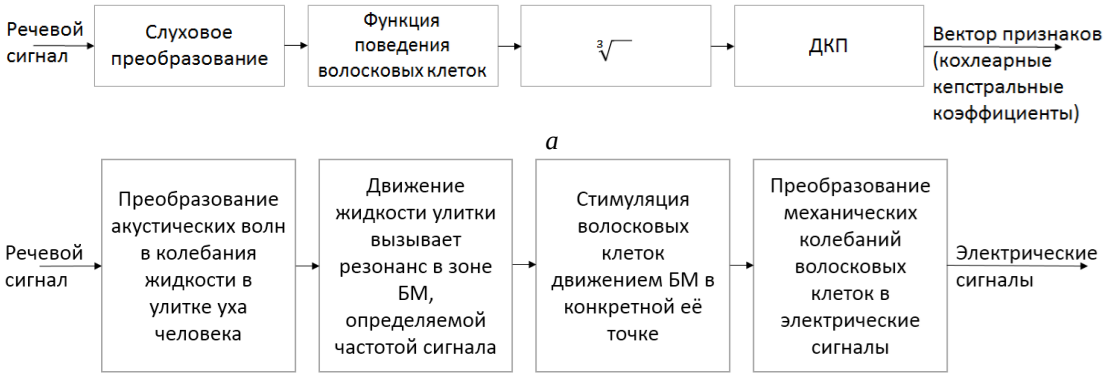


а *б*

Рис. 2. Преобразование на основе кохлеарной модели: *а* – импульсные характеристики; *б* – частотные характеристики банка фильтров

Этапы экстракции характеристического вектора с использованием преобразования речевого сигнала на основе кохлеарной модели

В работе [9] был предложен алгоритм получения характеристического вектора, который пытается смоделировать процессы обработки речевого сигнала человеческим ухом. Рассматриваемый алгоритм состоит из этапов, представленных на рис. 3, *а* и частично моделирует процесс обработки звукового сигнала слуховой системой человека.



а

б

Рис. 3. Сравнение этапов обработки сигнала: *а* – схема рассматриваемого алгоритма экстракции; *б* – процесс обработки звукового сигнала в слуховой системе человека

Результатом преобразования на основе кохлеарной модели является сигнал, разложенный на частотные полосы, так, как это делает базилярная мембрана в улитке уха человека. Далее по аналогии со слуховой системой человека моделируется поведение волосковых клеток (внутренних и внешних) и учет их нелинейности восприятия. После следует этап уменьшения размерности полученных характеристик. Полученный характеристический вектор называют кохлеарными кепстральными коэффициентами ККК (Cochlear Filter Cepstral Coefficients – CFCC) [8]. Функция поведения волосковых клеток была выбрана такая же, как в работе [8] – квадратичная функция. По данному алгоритму были получены характеристические признаки для дальнейшей идентификации диктора.

Конструирование системы идентификации диктора

В качестве решающих правил системы идентификации диктора наиболее часто используются следующие методы: векторное квантование, гауссовские смеси, нейронные сети и метод опорных векторов. В данном исследовании были выбраны нейронные сети прямого

распространения для получения сравнительных результатов распознавания диктора с результатами, представленными в работе [8], в которой были использованы модели гауссовых смесей. Для экспериментальной оценки системы идентификации диктора в акустических шумах был спроектирован идентификатор дикторов на основе нейронных сетей прямого распространения. Модель системы распознавания представлена на рис. 4 и имеет два режима работы: обучение и идентификация. Архитектура используемой сети выбиралась постепенно от простого однослойного персептрона к многослойным экспериментально. Используемые типы нейронных сетей: многослойный персептрон (два и три слоя); 2 слоя: 56 входных – 19 выходных. 3 слоя, 56 входных – 30 промежуточных – 19 выходных. Функция активации на всех слоях, кроме выходного: сигмоид, на выходном – функция sofmax. Процесс обучения происходит с учителем.

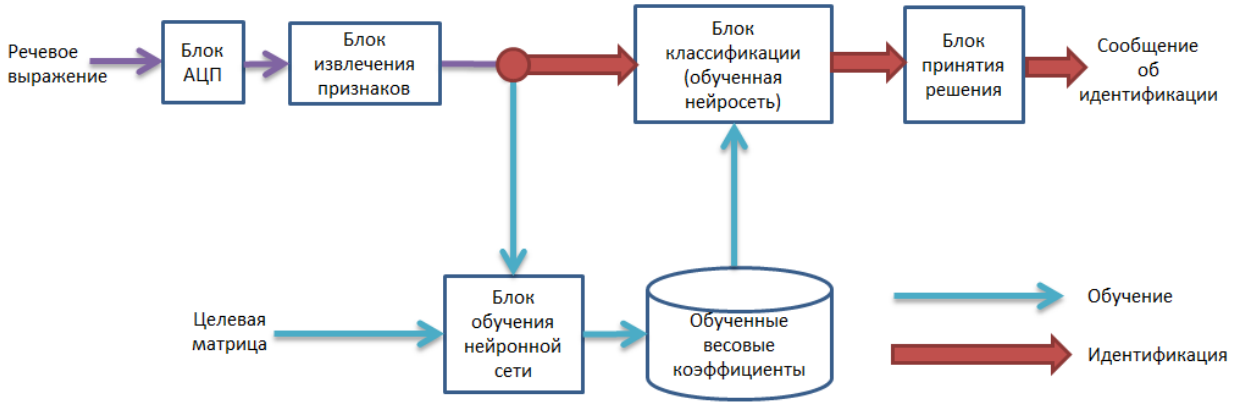


Рис. 4. Модель системы распознавания

Используемая речевая база была подготовлена на основе базы, сформированной для соревнований по разделению и распознаванию речи и дикторов в условиях с шумами [10]. Было выбрано 4 диктора, у каждого по 18 речевых фраз средней продолжительностью около 2 секунд, не содержащих шумы.

Оценка системы идентификации диктора в акустических шумах

В алгоритме экстракции характеристического вектора использовалось 32 кохлеарных фильтра [8]. Оценкой системы идентификации выступает точность распознавания диктора – отношение числа правильно распознанных речевых выражений к общему числу высказываний, участвовавших в распознавании. Данная оценка выражается в долях единицы или в процентах. Для получения сравнительных результатов был реализован алгоритм получения кепстральных коэффициентов в шкале Мел (мел-частотные кепстральные коэффициенты – МЧКК) и использован в системе идентификации диктора. Система идентификации реализована на языке Matlab. Результаты распознавания на обучающем множестве представлены на рис. 5. Тип используемого шума – лепет, бормотание (англ. «babble»).

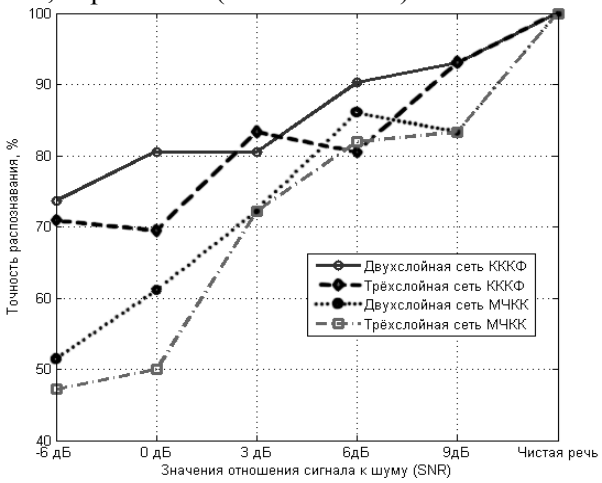


Рис. 5. Сравнение результатов для обучающего множества (тип шума – бормотание)

Как видно из рис. 6, в условиях без шумов признаки ККК показывают сравнимые результаты с признаками МЧКК и имеют 100 % точность распознавания (на ограниченном числе выражений). По мере того, как увеличивается уровень шума, точность системы распознавания падает, однако точность идентификации с использованием признаков ККК значительно лучше, чем МЧКК. Например, при SNR равным 0 дБ, точность распознавания для признаков МЧКК составляет 62 %, а для ККК – 80 %. На тестовом множестве точность распознавания всей системы снижается, однако результаты идентификации с использованием признаков ККК превосходят результаты МЧКК: 77 % к 69 % при SNR равным 0 дБ (рис. 6). Эффективность используемого алгоритма экстракции характеристического вектора в системе идентификации на нейронных сетях при уровне SNR –6 дБ больше на 20 % для МЧКК, и на 5–10 % для ККК, чем в системе на основе гауссовых смесей [8].

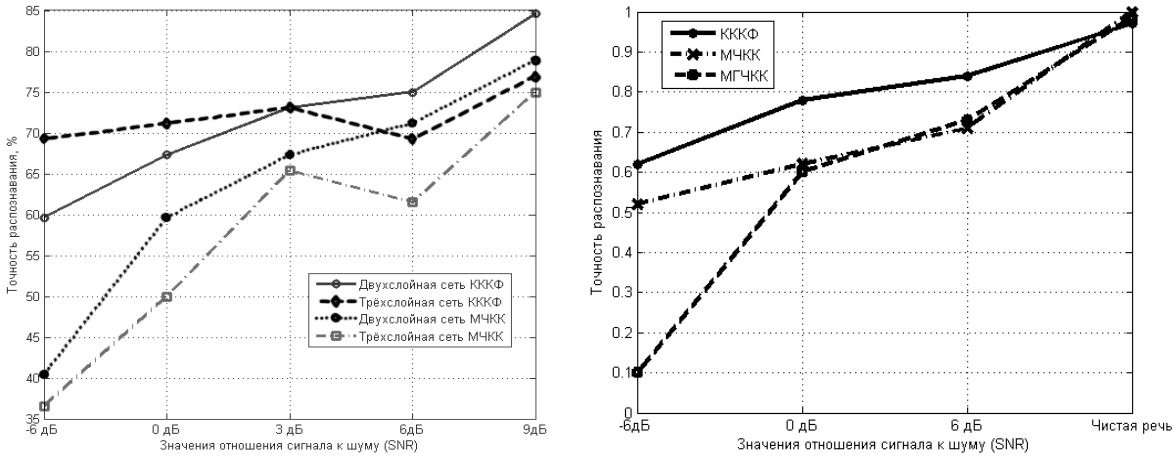


Рис. 6. Результаты распознавания для тестового множества (слева – система идентификации на нейронных сетях, справа – система идентификации на гауссовых смесях [8]) для типа шума – бормотание

Заключение

Разработана и реализована система идентификации диктора на основе психоакустически мотивированного алгоритма экстракции характеристического вектора с использованием нейронных сетей прямого распространения. Использование алгоритма извлечения признаков с преобразованием речевого сигнала на основе кохлеарной модели повышает эффективность распознавания диктора в акустических шумах, в сравнении с широко используемыми мел-частотными кепстральными коэффициентами, в среднем на 10–15 % при различных уровнях зашумления. Разработанная система идентификации диктора на основе нейронных сетей демонстрирует эффективность распознавания на уровне зашумления в 6 дБ на 10 % больше, а при остальных уровнях зашумления точность соизмерима с результатами идентификации системы на основе гауссовых смесей [8].

AN AUDITORY-BASED SPEAKER IDENTIFICATION SYSTEM IN ACOUSTIC NOISES

D.N. KRUCHOK, A.A. PETROVSKY

Abstract

The speaker recognition system in acoustic noises with the use of anthropomorphic speech signal processing is considered. A transform based on the cochlear model and its applications for speaker recognition are described in details. The characteristic vector obtained on the basis of this transform used as features for speaker identification. Neural network of direct distribution as decision rules are used. The results of the developed recognition systems of speaker identification are given.

Keywords: characteristic vector, cochlear model, speaker identification system.

Список литературы

1. *Сорокин В.Н., Вьюгин В.В., Тананыкин А.А.* // Информационные процессы. 2012. Т. 12, № 1. С. 1–30.
2. Информационное агентство. Бизнес. Новости компаний [Электронный ресурс].– Режим доступа: http://www.interfax.by/news/belarus/1196246. – Дата доступа : 17.12.2015.
3. *Togneri R., Pullella D.* // IEEE Circuits and systems magazine. 2011. P. 23–58.
4. *Mammone R., Zhang X., Ramachandran R.* // IEEE Signal processing magazine 1996. Vol. 13, № 5. P. 58–71.
5. Furui S. // IEEE Trans. Acoustics Speech Signal Process. 1981. Vol. 29, № 2. P. 254–272.
6. *Hermansky H., Morgan N.* // IEEE Trans. Speech Audio Process. 1994. Vol. 2, № 4. P. 578–589.
7. *Qi Li* // Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. New York, 2009. P. 181–184.
8. *Qi Li* // IEEE Transactions Audio, Speech, and Language Processing. 2011. Vol. 19, № 6. P. 1791–1801.
9. *Petrovsky A.A., Likhachov D.S., Wan W.* // Computing. 2004. Vol. 3, № 1. P. 75–83.
10. The PASCAL CHiME speech separation and recognition challenge [Electronic resource]. – Mode of access: http://spandh.dcs.shef.ac.uk/projects/chime/PCC/results.html. – Date of access : 19.01.2016.