

УДК 621.391.8

КОНВЕРСИЯ ПРОСОДИЧЕСКИХ ХАРАКТЕРИСТИК ДИКТОРА НА ОСНОВЕ МЕТОДОВ ПАРАМЕТРИЗАЦИИ КОНТУРА ЧАСТОТЫ ОСНОВНОГО ТОНА

В.А. ЗАХАРЬЕВ, А.А ПЕТРОВСКИЙ

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

Поступила в редакцию 26 июня 2012

Исследуется алгоритм конверсии просодических характеристик диктора на основе параметризации контура частоты основного тона. Рассмотрены основные этапы и методы, применяемые для построения алгоритма конверсии. Приведен сравнительный анализ предлагаемого подхода с конверсией частоты основного тона на основе метода нормализации по Гауссу.

Ключевые слова: конверсия голоса, просодические характеристики, параметризация контура частоты основного тона.

Введение

Конверсия голоса – это процесс преобразования параметров речевого сигнала, характеризующих одного диктора, в параметры другого, без изменения лингвистической составляющей самого сообщения. Первый диктор называется исходным, второй – целевым. Процесс конверсии подразумевает изменение акустических, фонетических и просодических характеристик исходного диктора в характеристики целевого согласно определенному набору правил, представляющему собой модель конверсии голоса. Перцептивное качество преобразования определяется точностью и сложностью построения модели конверсии, а также мерой того насколько хорошо могут быть аппроксимированы параметры исходного диктора параметрами целевого.

Системы конверсии голоса используются в широком кругу прикладных задач. Начиная от криминалистики, где они используются в качестве средств фонологической защиты свидетелей, заканчивая индустрией развлечений и системами мультимедиа. Характерным примером приложения в данной области является система «караоке», которая позволяет исполнителю, помимо песни, выбирать желаемый голос (например, известного певца), и в процессе работы в реальном времени преобразующая голос исполнителя в голос желаемой персоны. Отдельно необходимо отметить, что процесс внедрения технологии конверсии голоса имеет большую значимость для создания высококачественных систем синтеза речи по тексту. Он позволяет эффективно создавать мультимодальные (многоголосые), легко настраиваемые на произвольного диктора, системы синтеза, без значительных трудозатрат на создание речевых баз для каждого отдельного диктора.

В первом разделе статьи рассматривается общая схема построения системы конверсии голоса, а также раскрывается актуальность задачи конверсии просодических характеристик. Во втором разделе приводится описание стандартного алгоритма конверсии частоты основного тона (ЧОТ) по методу нормализации Гаусса, а также улучшенного, на основе параметризации контура ЧОТ, по модели Джиллета. В третьем разделе приводится алгоритм и даются пояснения методам, которые используются для определения параметров модели по Джиллету и синтеза сигнала с целевыми параметрами. В четвертом разделе приведены результаты сравнительного анализа двух, предложенных к рассмотрению, методов, сделаны выводы.

1. Конверсия просодических характеристик

Типовая обобщенная структурная схема системы конверсии голоса представлена на рис. 1 [1]. Процесс работы системы осуществляется в два этапа: обучения и конверсии.

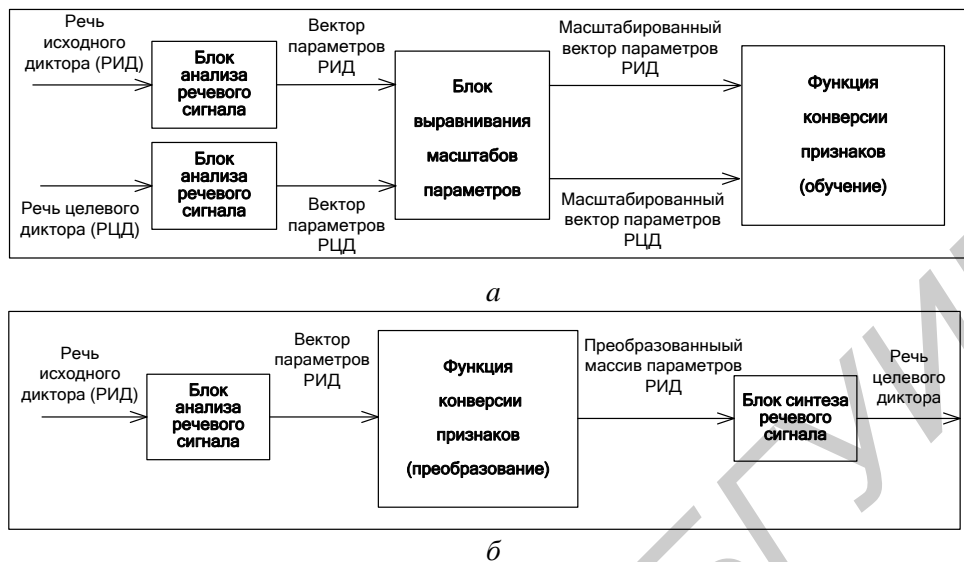


Рис 1. Структурная схема системы конверсии голоса:
а – на этапе обучения; б – на этапе конверсии

Этап обучения. На вход системы поступают речевые сигналы исходного и целевого дикторов. В соответствующем блоке производится анализ речевого сигнала, согласно некоторой модели представления, и отыскивается параметрическое представление для каждого фрейма входного сигнала. Для анализа сигнала может быть использован широкий спектр методов, как на основе моделей речеобразования (линейное предсказание, формантный и кепстральный анализ и т.д.), так и на основе сигнальных моделей (синусоидальный анализ, вейвлет анализ, мгновенный гармонический анализ [1] и т.д.). Далее для двух последовательностей векторов параметров производится операция масштабирования и перемещения друг относительно друга, для того, чтобы можно было выровнять их длину, а также установить временное соответствие. Затем начинается непосредственное формирование функции конверсии. С использованием алгоритма кластеризации «без учителя» (к-средних, Гауссовы смеси, деревья решений и т.д.) формируются дискретные или непрерывные подпространства пространства характеристических векторов исходного и целевого дикторов. Далее отыскивается функция конверсии, позволяющая найти, в оптимальном случае, однозначное отображение вектора параметров исходного диктора в параметры целевого. Этап обучения считается завершенным. Для каждой пары исходного и целевого дикторов задача обучения ставится только один раз.

Этап конверсии. На вход системы подается речевой сигнал только исходного диктора. Лингвистическая составляющая речевого сообщения, в общем случае, отличается от текста, произносимого на этапе обучения. Для каждого фрейма производится анализ речевого сигнала в соответствии с тем же методом, что и на этапе обучения. Далее осуществляется непосредственно процесс конверсии: согласно обученной модели конверсии, характеристический вектор исходного диктора преобразуется таким образом, чтобы максимально приблизиться к соответствующему характеристическому вектору целевого диктора. Полученная последовательность модифицированных векторов используется для синтеза речевого сигнала исходного диктора с характерными чертами целевого диктора.

Необходимо отметить, что при создании систем конверсии голоса разработчики уделяют основное внимание конверсии тембральных характеристик речи. Например, если руководствоваться моделью речеобразования источник-фильтр, внимание в основном уделяется конверсии полюсов передаточной функции, описывающих резонансные свойства речевого тракта. Однако речь является сложным многокомпонентным сигналом, содержащим в себе большое количество различной информации, в том числе о личности говорящего, которая мо-

жет выражаться в различных характеристиках речи. Выделяют артикуляторные и просодические характеристики. К артикуляторным можно отнести тембральные свойства голоса, задающие спектральную окраску фонем. Просодические характеристики – это совокупность физических параметров речевого сигнала, по средствам которых реализуются интонация и ударения в речи. Мелодика – это движение частоты основного тона. Ритмика – текущее изменение длительности звуков и пауз. Энергетика – текущее изменение силы звука [2]. В плане конверсии просодических характеристик внимание уделяется только мелодической составляющей просодики и, в основном, в рамках простейших моделей конверсии, речь о которых пойдет далее. Однако на данном этапе развития систем конверсии голоса становится очевидно, что использование более совершенных моделей конверсии просодических характеристик, дает возможность существенного улучшения параметров узнаваемости сконвертированного голоса.

2. Модели конверсии просодических характеристик

Основной моделью для конверсии мелодической составляющей просодических характеристик на данный момент является модель на основе метода нормализации Гаусса [3]. Функция конверсии представлена выражением

$$M_N(x) = \frac{(x - \mu_{src})}{\sigma_{src}} \cdot \sigma_{trg} + \mu_{trg}, \quad (1)$$

где x – значение частоты основного тона исходного диктора, $M_N(x)$ – значение частоты основного тона исходного диктора после конверсии, μ_{src} , σ_{src} – математическое ожидание и стандартное отклонение частоты основного тона исходного диктора, μ_{trg} , σ_{trg} – математическое ожидание и стандартное отклонение частоты основного тона целевого диктора.

Достоинствами такой модели конверсии являются простота реализации и низкая вычислительная сложность. Главным недостатком – то, что конверсия мелодики осуществляется на основе статистических характеристик, задающих самый общий вид изменения ЧОТ, что ведет к чрезмерному усреднению значений ЧОТ без учета локальных особенностей контура. Это, в свою очередь, приводит к деградации параметров естественности и узнаваемости системы конверсии голоса.

Второй моделью, рассмотрение реализации которой предлагается в данной работе, является модель, основанная на более детальной параметризации контура ЧОТ. Она основана на выделении в контуре частоты основного тона особых точек, согласно методике Паттерсона [4]. Пример фрагмента контура, а также особые точки контура представлены на рисунке.

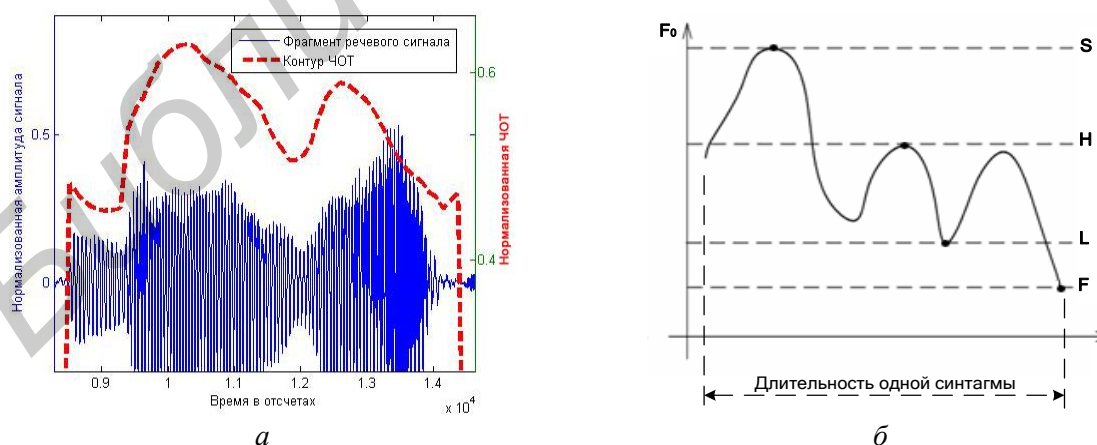


Рис 2. Параметризация контура частоты основного тона:

a – пример контура ЧОТ; b – особые точки контура согласно методике Паттерсона

Из рассмотрения рисунка следует, что характерными точками контура являются точки: S – точка, характеризующая максимальную высоту ЧОТ синтагмы [2], H – точка, определяющая акцентный пик ЧОТ внутри синтагмы, L – точка, характеризующая акцентный спад ЧОТ внутри синтагмы, F – точка, определяющая окончательный спад контура ЧОТ синтагмы. Параметр

S рассчитывается как глобальный максимум ЧОТ синтагмы, H как среднее значение локальных максимумов синтагмы, L как среднее значение локальных минимумов ЧОТ синтагмы, F как глобальный минимум ЧОТ синтагмы.

В процессе обучения модели конверсии для каждого диктора осуществляется поиск среднего значения особых точек по всей выборке обучающих фраз, которые и берутся в дальнейшем за значение конкретной точки. В конечном счете получается набор, состоящий из четырех обозначенных выше точек, которые должны хорошо аппроксимировать параметры мелодики для данного диктора.

Имея соответствующие наборы точек для исходного и целевого диктора, для конверсии просодических характеристик используется функция конверсии, предложенная Джиллетом [5], и представленная следующей формулой:

$$M_{PL}(x) = \begin{cases} F_{trg} + \frac{(x - F_{src})(L_{trg} - F_{trg})}{(L_{src} - F_{src})}, & x < L_{src}, \\ L_{trg} + \frac{(x - L_{src})(H_{trg} - L_{trg})}{(H_{src} - L_{src})}, & L_{src} \leq x \leq H_{src}, \\ H_{trg} + \frac{(x - H_{src})(S_{trg} - H_{trg})}{(S_{src} - H_{src})}, & x > H_{src}, \end{cases} \quad (2)$$

где x – значение частоты основного тона исходного диктора, $M_{PL}(x)$ – значение частоты основного тона исходного диктора после конверсии, $S_{src}, H_{src}, L_{src}, F_{src}$ – значение точек S, H, L, F для исходного диктора, $S_{trg}, H_{trg}, L_{trg}, F_{trg}$ – значение точек S, H, L, F для целевого диктора.

Анализируя выражение (2), легко заметить, что идейно оно весьма схоже с функцией конверсии по методу нормализации Гаусса, представленной выражением (1). Однако функция $M_{PL}(x)$ является кусочно-линейной, обладающей различными свойствами в различных областях значений частоты основного тона исходного диктора, что без значительного увеличения вычислительных затрат на расчет сконвертированного значения ЧОТ призвано обеспечить лучшее качество конверсии просодических характеристик.

3. Алгоритм работы системы конверсии просодических характеристик

Алгоритм работы системы конверсии просодических характеристик на основе модели Джиллета представлен на рис. 3. На этапе обучения он включает в себя следующие основные шаги.

Шаг 1. Подготовка сигнала. Фрагментация сигнала на фреймы, с перекрытием.

Шаг 2. Детектирование речевой активности по методу определения логарифма энергии текущего фрейма сигнала и подсчета меток пересечения нуля.

Шаг 3. Поиск контура частоты основного тона. Для поиска контура ЧОТ используется алгоритм определения ЧОТ на основе сопоставления сигнала с пилообразным импульсом (A sawtooth waveform inspired pitch estimator – SWIPE).

Шаг 4. Поиск локальных и глобальных точек экстремума контура ЧОТ для каждой из фраз.

Шаг 5. Расчет особых точек контура согласно методике Паттерсона, представленной во втором разделе. На этом процесс обучения считается завершенным.

Данный алгоритм используется для формирования множества характеристических точек контура ЧОТ как исходного, так и целевого диктора, соответственно приведенная выше последовательность действий выполняется два раза. Последовательно или параллельно будет идти этот процесс – не имеет значения, поскольку для определения параметров функции конверсии в данной системе нет необходимости во временном выравнивании и масштабировании контуров ЧОТ исходного и целевого диктора друг относительно друга.

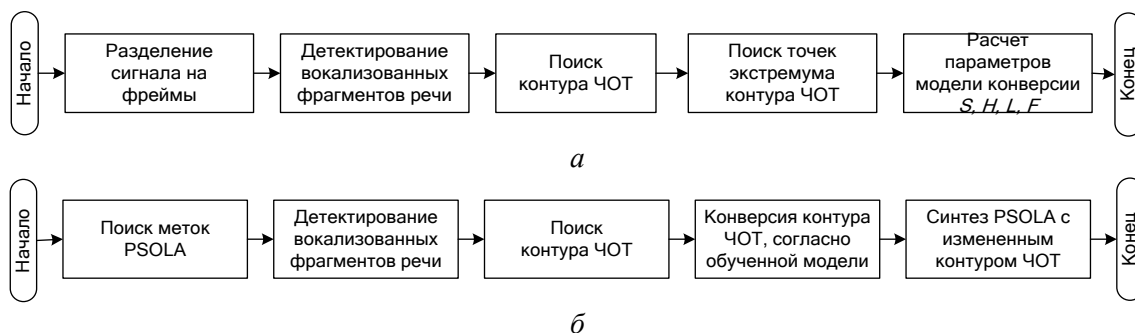


Рис 3. Алгоритм работы системы конверсии просодических характеристик:
a – на этапе обучения; *б* – на этапе конверсии

На этапе конверсии алгоритм включает в себя следующую последовательность действий.

Шаг 1. Автоматический поиск меток фундаментального периода (величина обратная ЧОТ) для дальнейшего их использования при синтезе сигнала, с перекрытием, синхронизированном с ЧОТ (Pitch Synchronous Overlap And Add – PSOLA). Иногда данный алгоритм называют PSOLA-анализом.

Данный шаг отличается от деления на фреймы, осуществляемого на этапе обучения, лишь тем, что сигнал фрагментируется не просто на равные части, а на отрезки, на которых период основного тона остается относительно постоянным.

Шаг 2. Детектирование вокализованных фрагментов. Процедура аналогична второму шагу на этапе обучения.

Шаг 3. Поиск контура частоты основного тона. Процедура аналогична третьему шагу на этапе обучения.

Шаг 4. Конверсия контура частоты основного тона в соответствии с моделью, представленной выражением (2).

Шаг 5. Синтез сигнала с использованием меток PSOLA, полученных на шаге один, с учетом сконвертированных значений контура ЧОТ шага четыре.

Для более полного пояснения хода процесса конверсии, остановимся на некоторых методах, задействованных в алгоритме поподробнее.

Детектирование вокализованных фрагментов речи. Поскольку наличие частоты основного тона характерно исключительно для вокализованных звуков речи, необходим механизм их определения в общем потоке. Для более точной идентификации разделение фреймов сигнала на вокализованную и невокализованную группы производится на основе пары методов, работающих вместе. Первый метод заключается в расчете логарифма энергии текущего фрейма сигнала, а также определении множества таких фреймов, для которых данное значение энергии превышает некоторое пороговое. Аналитическое выражение, позволяющее формализовать данный процесс, представлено выражением (3.1).

Второй метод – подсчет меток пересечения нуля. Основан на представлениях о том, что невокализованные и шумовые фрагменты речи имеют много большее количество пересечений функции сигнала с нулем, чем тональные его составляющие. Расчетная формула для него представлена выражением (3.2)

$$E(k) = \log \sum_{i=0}^{N-1} s(i)^2, \quad k = \overline{1 \div M},$$

$$E_{\text{flag}}(k) = \begin{cases} 1, & E(k) < E_{\text{THR}} \\ 0, & E(k) \geq E_{\text{THR}} \end{cases}, \quad (3.1)$$

$$Z(k) = \sum_{i=1}^{N-1} \text{abs}(\text{sign}(s(i)) - \text{sign}(s(i-1))),$$

$$Z(k) = \frac{Z(k)}{2 \cdot N}, \quad k = \overline{1 \div M}, \quad (3.2)$$

$$Z_{\text{flag}}(k) = \begin{cases} 1, & Z(k) < Z_{\text{THR}} \\ 0, & Z(k) \geq Z_{\text{THR}} \end{cases},$$

где N – количество отсчетов в одном фрейме сигнала, M – количество фреймов, $s(i)$ – амплитуда i -го отсчета, $E(k)$ – значение энергии сигнала k -го фрейма, $E_{\text{flag}}(k)$ – отметка о вокализованности k -го фрейма на основе метода логарифма энергии, $Z(k)$ – количество пересечений отметок нуля на k -ом фрейме, $Z_{\text{flag}}(k)$ – отметка о вокализованности k -го фрейма на основе метода отметок пересечения нуля, $E_{\text{THR}}, Z_{\text{THR}} = \text{const}$ – пороговые значения энергии и количества пересечений нуля, соответственно.

Поиск контура частоты основного тона. Ведется на основе алгоритма SWIPE [6]. Данный алгоритм позволяет оценить ЧОТ как фундаментальную частоту пилообразного сигнала, чей спектр наиболее точно соответствует спектру исследуемого входного сигнала. Последовательность действий состоит из двух основных шагов. На первом определяются кандидаты ЧОТ – для каждого из них в диапазоне поиска ЧОТ вычисляется его интенсивность. Интенсивность тем выше, чем больше степень соответствия спектра сигнала на основе кандидата спектру пилообразного сигнала. На втором этапе ЧОТ определяется как частота кандидата с наибольшей интенсивностью. Основные его достоинства – то, что в методе учитываются психоакустические особенности восприятия человеком различных частотных компонент, поиск ведется в перцептуально-мотивированной частотной шкале, используются понятия эквивалентных прямоугольных полос, а также частотного маскирования. Среди всех протестированных методов оценки, SWIPE позволяет получить наиболее четкий контур ЧОТ с возможностью точной локализации его особенностей.

Анализ и синтез сигнала с перекрытием, синхронизированный с ЧОТ. Для модификации контура ЧОТ на этапе используется метод PSOLA [7]. Алгоритм включает в себя следующие основные шаги. Определение разметки маркеров PSOLA (рис. 4, а). Их расположение должно удовлетворять двум основным условиям: расстояние между соседними метками должно быть близким к фундаментальному периоду; маркеры должны располагаться в местах локального максимума энергии. Декомпозиция сигнала на сегменты производится относительно разметки.

Модификация разметки – расстояния между маркерами – согласно сконвертированным значениям меток контура ЧОТ, получившимся на выходе модели конверсии (рис. 4, б). Последним шагом является процесс синтеза сигнала методом перекрытия со сложением.

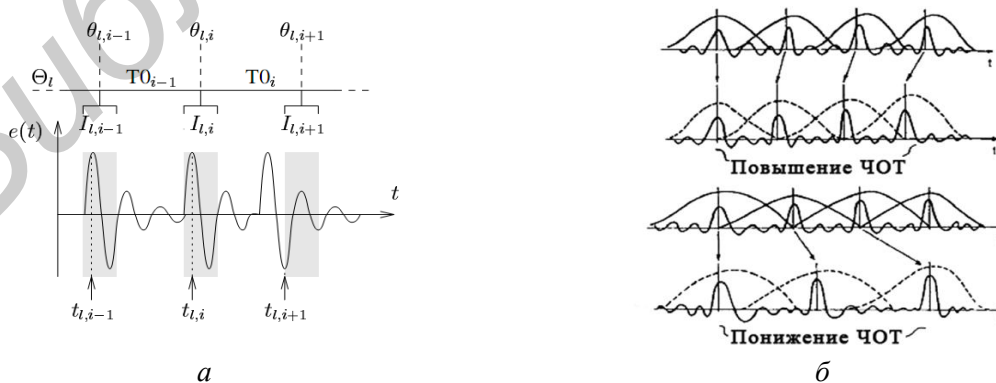


Рис 4. Анализ-синтез сигнала по методу перекрытия со сложением синхронизированный с ЧОТ: а – процесс расстановки меток PSOLA в сигнале; б – модификация

4. Сравнительный анализ методов конверсии

Эксперименты по эффективности использования предложенного метода проводились на следующей выборке речевых сигналов. Для обучения моделей конверсии была использована речевая база фонетически сбалансированных предложений Санкт-Петербургского лингвистического университета. Были выбраны записи речевых сигналов двух славяноговорящих дикторов разных полов длительностью 4 мин. Частота дискретизации звукозаписи – 16000 Гц, разрядность – 16 бит. Лингвистическое содержание звукозаписей идентичное. Для обучения использовалась первая минута одной и второй фонограмм, для тестирования – фразы различной длительности, в оставшемся 3-минутном фрагменте фонограммы. Результаты сравнительного анализа работы двух методов конверсии просодических характеристик представлены в таблице. Оценки качества определены на основе метода средней оценки мнений экспертов (Mean Opinions Scores – MOS) [8].

Результаты экспериментов по конверсии просодических характеристик

Модель конверсии голоса	Критерии оценки			
	Разборчивость		Узнаваемость	
	мужского в женский	женского в мужской	мужского в женский	женского в мужской
Нормализация по Гауссу	4	3,5	4	3,5
Модель Джиллета	3,5	3,5	4,5	4,5

Из таблицы видно, что при использовании модели конверсии Джиллета можно получить большую степень соответствия (узнаваемости) мелодической составляющей просодических характеристик исходного диктора целевому.

Заключение

В данной статье были исследованы методы конверсии просодических характеристик на основе параметризации контура ЧОТ. Представлена модель на базе метода нормализации контура ЧОТ по Гауссу, и модель Джиллета. Описан алгоритм работы системы на основе данной модели. В ходе экспериментов показано, что представленный метод имеет большую перспективу для применения в системах конверсии с точки зрения критерия узнаваемости.

CONVERSION OF PROSODIC CHARACTERISTICS BASED ON THE PITCH CONTOUR PARAMETRIZATION

V.A. ZAKHARYEU, A.A. PETROVSKY

Abstract

A method of prosodic characteristics conversion based on the pitch contour parameterization, beside Patterson special points and Gillet conversion model, is presented in this paper. The comparison analysis showed better quality of proposed method versus canonical method of pitch Gaussian normalization. Therefore, it has a greater future for use in the conversion systems from the standpoint of the criterion of recognition.

Список литературы

1. Петровский А.А. Анализаторы речевых и звуковых сигналов: методы, алгоритмы и практика. Минск, 2009.
2. Лобанов Б.М. Компьютерный синтез и клонирование речи. Минск, 2008.
3. Arslan M., Talkin D. // Proc. Eurospeech. 1997. P. 1347–1350.
4. Patterson D. A linguistic approach to Pitch Range Modelling. Edinburg, 2000.
5. Gillet B., King S. // Eurospeech. 2003. P. 101–104.
6. Camacho A. A sawtooth waveform inspired pitch estimator for speech and music. Florida, 2007.
7. Gold B., Morgan N. Speech and audio signal processing. Danvers, 2000.
8. Methods of subjective determination of transmission quality. ITU-T. 1996.