УДК 004.934.2

АКУСТИЧЕСКИЙ АНАЛИЗ РЕЧИ ЛИЦ С ДИСФОНИЕЙ

С.М. ГОРОШКО, канд. техн. наук, доц. С.Н. ПЕТРОВ
(Белорусский государственный университет информатики и радиоэлектроники)

Показана возможность при использовании акустического анализа извлекать параметры речи, которые могут быть использованы для диагностики голосовых патологий. Предлагается методика распознавания речи с использованием мел-частотных кепстральных коэффициентов. Выявление расстройств голосового аппарата предлагается свести к решению задачи кластерного анализа, в которой для количественной оценки сходства или различия между классифицируемыми объектами используется Евклидово расстояние между полученным вектором признаков речевого сигнала и вектором признаков набора тестовых данных. Методика позволяет определить различие между здоровым человеком и человеком с нарушением речевой функции и может быть использована для разработки устройств, предназначенных для диагностики дисфонии у лиц голосоречевых профессий.

Ключевые слова: *выявление патологии речи, распознавание речи, дисфония, мел-частотные кепстральные коэффициенты, дискретное преобразование Фурье, Евклидово расстояние.*

Введение. В настоящее время возросло число лиц голосоречевых профессий. Таким людям в своей профессиональной деятельности необходимо обладать голосом достаточной силы и интонационной выразительности для достижения убедительности, так как смысл высказывания во многом определяется не только лексикой, но и интонацией. Акустические характеристики голоса оказывают значительное влияние на интонацию, артикуляцию и дикцию.

Речь – это сложный психофизический процесс, результат работы мышц, связок, дыхания, центральной и периферической нервной системы. Нарушения голосовой функции объединяют под общим названием «дисфония». Подобные нарушения могут носить врожденный или приобретенный характер в результате травм, операций, нервно-психических, инфекционных или иных заболеваний. Большинство заболеваний гортани и патологии голосовых связок вызывают значительные изменения в речи. Также известно, что на голос в значительной степени влияет текущее психофизическое состояние человека [1].

Хроническая дисфония возникает при органических повреждениях голосовых связок, таких как полип, отек Рейнке и т.д., а также при раке горла, неврологических расстройствах и хроническом раздражении, вызванном курением.

В число традиционных процедур для осмотра гортани с клиническими или исследовательскими целями входят гибкая и жесткая фиброскопическая ларингоскопия (обследование с помощью волоконно-оптического инструмента), видеостробоскопия (стробовое освещение гортани, полезное для визуализации движений), электромиография (косвенное наблюдение за функциональным состоянием гортани) и видеофлуороскопия (рентгенографическая техника, при которой пациент глотает радиоактивное непрозрачное вещество для оценки функции глотания). Все они имеют инвазивный характер.

В последние годы возрос интерес к акустическому анализу голоса как альтернативному средству диагностики. Этот тип анализа обладает преимуществами перед существующими методами обследования из-за его неинвазивного характера, а также возможности предоставить количественные данные о клиническом состоянии функций гортани и голосового тракта с соответствующими результатами временного анализа. В частности, акустический анализ находит широкое применение для обследования детей с задержкой или нарушениями речевого развития.

Получение вектора признаков на основе мел-частотных кепстральных коэффициентов. Основная суть данной методики заключается в получении основных характеристик речевого сигнала и отбрасывании индивидуальных особенностей произношения слов диктором. Преимущество кепстральных коэффициентов (Mel-frequency cepstral coefficients – MFCC) перед другими методами заключается в простоте реализации при достаточно высоком качестве распознавания речи. Мел – психофизическая единица высоты звука [2]. Данная оценка по высоте определена эмпирически, методом статистической обработки больших сведений о восприятии звука человеком. Высота звука логарифмически зависит от частоты звука. Использование мел-шкалы хорошо работает для чистого тона, при наличии посторонних шумов эффективность данной оценки сильно падает.

У каждого человека основные частоты и частоты формант индивидуальны, поэтому необходимо найти параметр, который учитывал бы эту особенность. Используя метод связывания с гармониками, можно избавиться от любых фиксированных значений анализируемых полос частот и наблюдать изменения характеристик речи в динамике [3]. Однако есть сложность в описании параметров каждой фор-

манты отдельно, поскольку они не всегда явно выражены. Поэтому основным параметром в методах анализа сигнала должна считаться основная частота голоса человека. Отношение интегрального спектра высокочастотной полосы к спектру полосы низкой частоты с привязкой к гармоникам является наиболее эффективным методом определения речевых патологий. Изучение результатов проведенных исследований показало целесообразность использования предложенного метода для анализа нарушений речи [4].

Для этого использовался речевой сигнал, полученный у диктора с использованием микрофона. Блок-схема алгоритма обработки речевого сигнала представлена на рисунке 1.

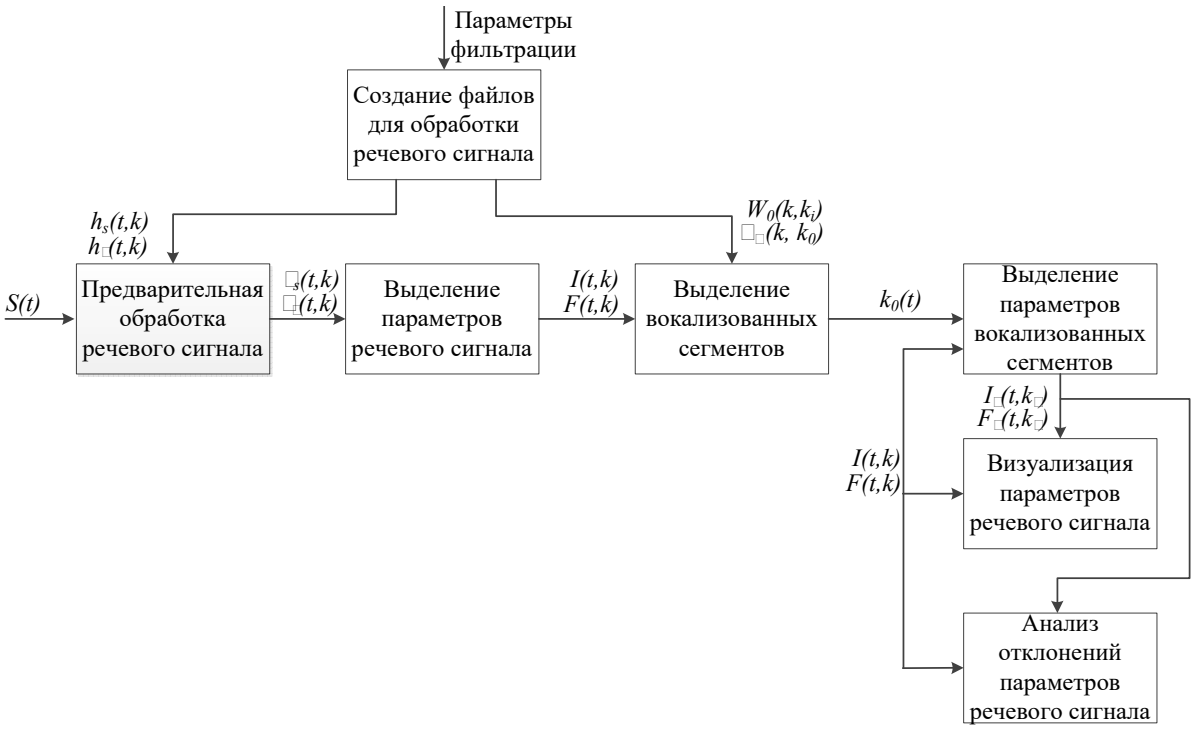


Рисунок 1. – Блок-схема алгоритма обработки речевого сигнала

Речевой сигнал имеет конечную длительность и не является периодическим, поэтому необходимо разбить этот сигнал на сегменты (кадры) конечной длины, спектры которых остаются относительно неизменными в течение выбранного периода времени. Обычно за такой период принимается длительность 5–100 мс. Разбиение исходного сигнала на кадры заданной длительности делается с половинным перекрытием для борьбы с искажениями, которые могут быть вызваны расположенными рядом кадрами, либо без перекрытия, что значительно экономит вычислительные ресурсы. В пределах полученного кадра входной сигнал можно считать стационарным. С точки зрения динамики речи самые быстрые изменения могут происходить всего за несколько миллисекунд, тогда как некоторые гласные звуки остаются относительно стабильными в течение 100–200 мс. Часто используются кадры длиной 10–25 мс [5].

В данной работе использовались кадры длительностью 10 мс с половинным наложением (5 мс).

Кроме того, вследствие конечной длительности и непериодического характера речевого сигнала из-за разрывов на концах временного ряда при применении преобразования Фурье проявляется эффект утечки. Эффект спектральной утечки затрудняет корректное выделение гармонических составляющих, близких друг к другу по частоте, из-за конечной ширины главного лепестка, а также вносит искажения в далеко расположенные частоты, мешая точному измерению амплитуды отдельных частот в связи с наличием боковых лепестков.

Для снижения влияния эффекта утечки на конечный результат применялась весовая функция Хэмминга, что позволило минимизировать максимумы боковых лепестков:

$$w(n) = 0,54 - 0,46\cos\left(\frac{2\pi n}{N-1}\right),\ n = 0,\ldots,N-1, \qquad (1)$$

где N – длина кадра.

Далее для каждого кадра вычислось дискретное преобразование Фурье (ДПФ). ДПФ взвешенного сигнала имеет вид

$$X_k = \sum_{n=0}^{N-1} x_n w_n \exp(-\frac{2\pi i}{N} kn), \quad (2)$$

где k – частоты.

$$f_k = \frac{F_s}{N} k, \quad (3)$$

где F_s – частота дискретизации сигнала.

Следующим шагом вычисляем периодограмму – оценку спектральной плотности мощности (СПМ)

$$P_i(k) = \frac{|Y_i(k)|^2}{N}. \quad (4)$$

Полученное представление сигнала в частотной области разбивают на диапазоны с помощью набора (банка) из 26 треугольных фильтров. Границы фильтров рассчитывают в шкале мел. Переход от частотной шкалы к мел-шкале осуществляется по формуле

$$M(f) = 1127 \ln\left(1 + \frac{f}{700}\right), \quad (5)$$

где $M(f)$ – частота в мелах.

Переход обратно к частотной шкале осуществляется по формуле

$$F(m) = 700\left(\exp\left(\frac{m}{1125}\right) - 1\right). \quad (6)$$

Фильтры умножаются на периодограмму кадра и вычисляется энергия для каждого фильтра (по 26 для каждого кадра). Полученные энергии логарифмируются:

$$e_m = \ln\left(\sum_{k=0}^{N} |X_k|^2 H_{m.k}\right), \ m = 0, \ldots, N_{FB} - 1, \quad (7)$$

где N_{FB} – количество фильтров;

$H_{m.k}$ – весовые коэффициенты полученных фильтров.

На заключительном этапе расчета MFCC применяется дискретное косинусное преобразование. На выходе получаем многомерный вектор признаков сигнала, которые и являются мел-частотными кепстральными коэффициентами:

$$c_i = \sum_{m=0}^{N_{FB}-1} e_m \cos\left(\frac{\pi i(m + 0{,}5)}{N_{FB}}\right), \ i = 1, \ldots, N_{MFCC}. \quad (8)$$

Коэффициент c_0 не используется, так как представляет собой энергию сигнала.

Методика проведения эксперимента. В качестве эталонных использовались фразы, приведенные в приложение Д ГОСТ Р 50840-95. Тестовые фразы были записаны с использованием микрофона AKG P120, подключенного к аудиоинтерфейсу Focusrite Scarlett 2i2. Данные сохранялись и обрабатывались на персональном компьютере. Чтение осуществлялось дикторами ровным голосом, четко, но без подчеркивания отдельных звуков с постоянным уровнем речи. На протяжении чтения всей таблицы дикторами выдерживался постоянный ритм речи [6]. Предполагается создание нескольких тестовых баз. Одна из них создается из записей лиц без патологий речи, другая (или другие) создается из записей людей с дисфонией.

Далее проводится сравнение акустических параметров исходных тестовых фраз и параметров фраз, полученных при произношении лицом, проходящим диагностику. Во время проведения диагностики, дикторам было предложено прочитать в общей сложности 80 различных фраз, в которых представлены наиболее значимые гласные в виде ударных фонем. При этом возможно оценить работу голосовых складок.

Задачу выявления расстройства голосового аппарата предлагается свести к задаче кластерного анализа, при которой для количественной оценки сходства или различия между классифицируемыми объектами используется понятие метрики. В качестве такой метрики использовалось Евклидово расстояние между полученным вектором признаков речевого сигнала и вектором признаков сигнала набора тестовых данных. Евклидово расстояние рассчитывается по формуле

$$D_e(a,b) = \sum_{i=1}^{N} \sqrt{(b_i^2 - a_i^2)}, \tag{9}$$

где D_e – расстояние между объектами b и a;

a_i – значение i-свойства объекта a;

b_i – значение i-свойства объекта b.

Необходимо провести кластеризацию признаков с одним или несколькими центрами для здоровых людей и людей с дисфонией. В результате чего получим 2 кластера. Вычислить значения вектора для лица, проходящего обследование. Найти расстояние от этого вектора до центров тестовых кластеров. В зависимости от того, к какому из кластеров окажется ближе полученный «диагностический» вектор, принимается решение наличии и об отсутствии расстройства речевой функции.

Реализация обработки данных средствами пакета MatLab 2017.

Для чтения звуковых файлов в среде разработки Matlab предусмотрена функция [sig, fs, b] = wavread (‘filename.wav’). Функция wavread загружает звуковой файл и возвращает выбранные данные: sig – звуковой сигнал со значениями от –1 до 1, fs – частота дискретизации, b – количество бит на одну выборку.

Разбиение записанного речевого сигнала на фреймы производится с использованием функции enframe (sig, win, inc), где win – оконная функция (Хэмминга), inc – смещение, а также для умножения на оконную функцию (формула 1).

Расчет преобразования Фурье в действительной области в среде Matlab производится функцией rfft (sig), где sig – умноженный на оконную функцию сигнал.

Функция melbank (p, n, fs, flow, fhigh, w) создает набор мел-фильтров, где p – количество фильтров, n – длина преобразования Фурье, fs – частота дискретизации, flow/fhigh – нижняя/верхняя границы фильтров, w – тип фильтров.

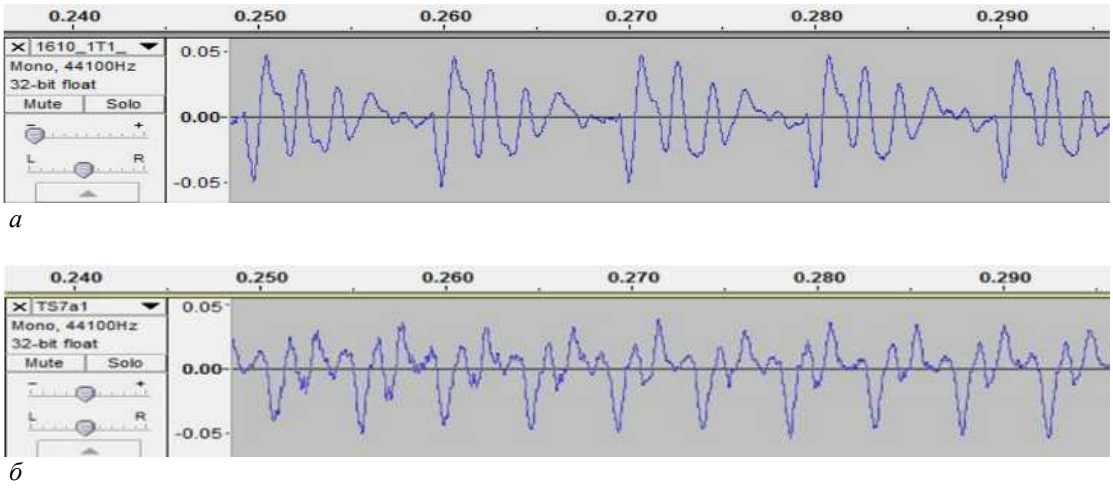
Функция rdct (x) реализует дискретное косинусное преобразование.

Функция melcepst (s, fs, w, nc, p, n, inc, flow, fhigh) возвращает мел-кепстральные коэффициенты, используя вышеназванные функции [7].

Рисунок 2. – Алгоритм вычисления мел-частотных кепстральных коэффициентов в Matlab 2017

Акустические характеристики речи при дисфонии. На рисунке 3 приведены фрагменты звукозаписи речи здорового человека (*а*) и человека с дисфонией (*б*). Запись голоса выполнялась в тихой комнате с микрофоном, расположенным на расстоянии 10 см ото рта. Проведение записи происходило с частотой дискретизации 44,1 кГц.

В качестве дикторов были приглашены преподаватели БГУИР, в том числе имеющие в анамнезе диагноз дисфонии.

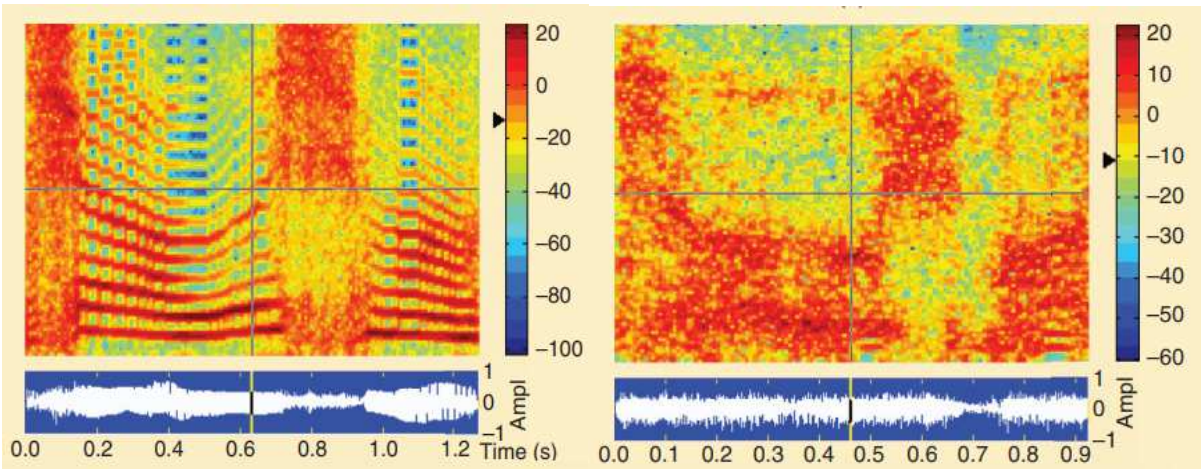


а

б

Рисунок 3. – Фрагмент звукозаписи речевого сигнала:
***а* – здоровый человек; *б* – человек с дисфонией**

Такой голос имеет более низкую амплитуду и по своим характеристикам напоминает шепот. Речь с дисфонией отличается от нормальной по звучанию тональностью и формантной структурой. На рисунке 4 приведены спектрограммы здорового человека и человека с дисфонией.



а *б*

Рисунок 4. – Спектрограмма речи:
***а* –здорового человека; *б* – человека с дисфонией**

На рисунке 4 отчетливо видно, что в сравнении с нормальной речью в вокализованных фонемах человека с дисфонией не наблюдается период основного тона. В дополнение к этому вокализованные фонемы отличаются от озвученных фонем нормальной речи с точки зрения формантного искажения. Ширина полосы дисфонических фонем больше, а их формантные частоты выше. Однако в невокализованных фонемах дисфонической речи не наблюдается значительных формантных искажений. Сравнивая дисфоническую речь с нормальной с точки зрения характеристик тональности, озвучивания и формантного искажения, было установлено, что для невокализованных фонем дисфонической речи не должно быть никаких изменений.

Заключение. В данной работе представлена система ранней диагностики лиц с дисфонией. В ее основу положены обнаружение и сравнение мел-частотных кепстральных коэффициентов речи, а также кластерный анализ. Методика позволяет дополнить традиционные инвазивные методы. В дальнейшем может быть расширена для диагностики большего числа нарушений речи, например, таких как дизарт-

рия, определения степени тяжести заболеваний или вида расстройств (гипотонусная, гипертонусная или гипо-гипертонусная дисфония).

ЛИТЕРАТУРА

1. Фомина, М.В. Профилактика и лечение нарушений голоса у лиц голосоречевых профессий : метод. указания / М.В. Фомина. – Оренбург : ГОУ ОГУ, 2004. – 19 с.
2. Воробьева, С.А. Выделение границ фонем речевого сигнала с помощью мел-частотных спектральных коэффициентов / С.А. Воробьева // Молодой ученый. – 2017. – № 13. – С. 2–6.
3. Малинин, П.В. Иерархический подход в задаче идинтификации личности по голосу с помощью проекционных методов классификации многомерных данных / П.В. Малинин, В.В. Поляков // Доклады Томского гос. ун-та систем управления и радиоэлектроники. – 2010. – № 1/1.
4. Первушин, Е.А. Обзор основных методов распознавания дикторов / Е.А. Первушин // Математические структуры и моделирование. – 2011. – Вып. 24.
5. Сорокин, В.Н. Распознавание личности по голосу: аналитический обзор / В.Н. Сорокин, В.В. Вьюгин, А.А. Тананыкин // Информационные процессы. – 2012. – Т. 12, №1.
6. Передача речи по трактам связи. Методы оценки качества, разборчивости, узнаваемости : ГОСТ Р 50840-95. – М. : Госстандарт России.
7. Смит С. Цифровая обработка сигналов. Практическое руководство для инженеров и научных работников / С. Смит. – М. : Додэка-XXI, 2012. – 720 с.

Поступила 27.02.2019

ACOUSTICAL ANALYSIS OF THE SPEECH OF PERSONS WITH DYSFONIA

S. GOROSHCKO, S. PETROV

The possibility of using acoustic analysis to extract speech parameters that can be used for the diagnosis of voice disability is shown. The method of speech recognition using Mel-frequency cepstral coefficients is proposed. Detection of voice disorders is reduced to solving a cluster analysis problem, in which the Euclidean distance between the received speech feature vector and the feature vector of the test data set is used to quantify the similarities or differences between the classified objects. The technique allows to determine the difference between a healthy person and a person with impaired speech function and can be used to develop devices designed to diagnose dysphonia in persons of voice-speech professions.

Keywords: *detection of speech pathology, speech recognition, dysphonia, mel-frequency cepstral coefficients, discrete Fourier transform, Euclidean distance.*