ТЕХНИЧЕСКИЕ НАУКИ

УДК 004.896 Научная статья
DOI: 10.35330/1991-6639-2022-1-105-41-58

АНАЛИТИЧЕСКИЙ ОБЗОР И КЛАССИФИКАЦИЯ МЕТОДОВ ВЫДЕЛЕНИЯ ПРИЗНАКОВ АКУСТИЧЕСКОГО СИГНАЛА В РЕЧЕВЫХ СИСТЕМАХ

И.А. ГУРТУЕВА, К.Ч. БЖИХАТЛОВ

Институт информатики и проблем регионального управления –
филиал Кабардино-Балкарского научного центра Российской академии наук
360000, Россия, Нальчик, ул. И. Арманд, 37-а

Аннотация. В данной работе представлен обзор методов и алгоритмов выделения признаков при трансформации акустического сигнала в последовательность векторов для решения задач сегментации, классификации, идентификации или распознавания речи. Предложена классификация методов извлечения признаков по математическим подходам. Обсуждаются алгоритмы и техники спектрального анализа, наиболее применяемые при проектировании систем распознавания речи. Настоящий обзор наглядно демонстрирует сложность проблемы акустической обработки – отыскания представления, снижающего размерность модели при сохранении полноты лингвистической информации и, что важно, устойчивого к вариативности относительно диктора, каналов передачи и окружающей среды. Проведенный анализ существующих методов извлечения признаков полезен для выбора технологии при проектировании ключевого элемента речевой системы.

Ключевые слова: распознавание речи, преобразование Фурье, кепстральный анализ, линейное предсказание, методы выделения признаков

Статья поступила в редакцию 08.02.2022 *Принята к публикации 14.02.2022*

Для цитирования. Гуртуева И.А., Бжихатлов К.Ч. Аналитический обзор и классификация методов выделения признаков акустического сигнала в речевых системах // Известия Кабардино-Балкарского научного центра РАН. 2022. № 1 (105). С. 41–58. DOI: 10.35330/1991-6639-2022-1-105-41-58

1. Введение

В настоящее время понятие «автоматическое распознавание речи» охватывает широкую сферу научной и инженерной деятельности. Речевые технологии находят применение в самых разных областях человеческой деятельности: от систем естественно-языкового управления программным и аппаратным обеспечением до использования голосовых ключей и помощи людям с ограниченными возможностями здоровья (глухие, слабослышащие, с травмами рук). Несмотря на столь разные назначения, при техническом проектировании речевых систем используется практически одна и та же процедура обработки звукового сообщения с незначительными вариациями. В большинстве механизмов автоматического распознавания речи могут быть выделены следующие модули: сбор данных, предварительная обработка, подсистема выделения полезного сигнала, блоки акустического и языкового моделирования и, наконец, декодирование.

Модуль сбора данных, часто называемый формированием спектра, конвертирует «сырой» аналоговый сигнал в форму, удобную для последующей обработки. Формирование спектра нацелено на создание дискретизированного представления речевых данных с наиболее высоким соотношением сигнала к шуму [1].

Подсистема предварительной обработки сигнала нацелена на исключение влияния окружающей среды, улучшение качества полезного сигнала [2], часто включает в себя подавление эха и усиление.

Блок выделения признаков преобразует сигнал в набор акустических параметров. Речь трансформируется для выделения релевантных характеристик и сжимается для упрощения последующей обработки. Техники сжатия размерности классифицируются по двум большим группам – извлечение и отбор признаков. Экстрагирование признаков – это интеллектуальная обработка, нацеленная на выявление знаний, неявным образом присутствующих в обрабатываемой информации. Это подход, при котором признаки проецируются в новое признаковое пространство меньшей размерности [3]. Отбор признаков также нацелен на выбор меньшего по объему набора признаков, который минимизирует избыточность и способствует максимальному соответствию задачам конкретного речевого приложения, однако использует исходные значения признаков в сжатом пространстве. Оба подхода для снижения размерности нацелены на оптимизацию процесса обучения, снижение сложности вычислений, построение лучше обобщающих моделей и экономии памяти.

Далее на стадии акустического моделирования осуществляется акустический анализ [2, 4] – определение степени подобия анализируемого отрезка речи и эталонов внутренней библиотеки системы. Акустический анализ осуществляется наложением каждой акустической модели на каждый речевой фрейм; на выходе получается матрица оценки фреймов. Существует большое число акустических моделей, отличных друг от друга по их представлению, зернистости, контекстной зависимости и другим свойствам. Оценки вычисляются в зависимости от типа использованной акустической модели. Для акустических моделей, основанных на шаблонном подходе, оценки представляют собой Евклидово расстояние между фреймом шаблона и фреймом неизвестного сигнала. Для акустических моделей на состояниях оценка представляет собой эмиссионную вероятность, то есть сходство текущего состояния, генерирующего текущий фрейм согласно параметрической или непараметрической функции состояния.

Оценка фреймов конвертируется в последовательность слов на основе идентификации последовательности акустических моделей, представляя собой значимую последовательность слов, которая дает наилучшую суммарную оценку вдоль пути выравнивания по матрице. Конечным результатом выравнивания по времени является последовательность слов как гипотеза о предложении, соответствующем распознаваемому высказыванию. На практике обычно возвращается несколько таких предложений с наивысшими оценками с использованием варианта выравнивания по времени, называемого *N-best search* [4].

Эффективность работы каждого последующего компонента и системы распознавания в целом зависит от двух ключевых подсистем: обработки сигнала и извлечения фичей. Практика выявила, что проблема акустической обработки, то есть решение о том, какие акустические данные будут оцениваться, нетривиальна. Найти представление, которое снижает сложность модели, сохранив при этом полноту лингвистической информации, несмотря на эффекты, возникающие как следствие вариативности диктора, технических средств передачи, окружающей среды.

2. Анализ методов и алгоритмов выделения акустических признаков

Блок спектрального анализа предназначен для трансформации речевого сообщения из акустической формы в набор информативных параметров, способных с достаточной полнотой дифференцировать звуковой образ. Целью данного этапа обработки является избавление от нерелевантных и избыточных для данного типа речевых приложений признаков и детализация релевантной информации. Выходом является неоднородный параметриче-

ский вектор, сочетающий в своем составе абсолютные и динамические (дискретные производные, чаще всего первые и вторые) спектральные признаки.

К настоящему времени сформировалось девять классов алгоритмов акустического анализа для последующего применения в автоматическом распознавании речи, идентификации языка, идентификации/верификации диктора и т. д. На рисунке 1 показаны три подхода, разработанные для решения задачи акустического анализа и семейства алгоритмов, созданных на их основе.

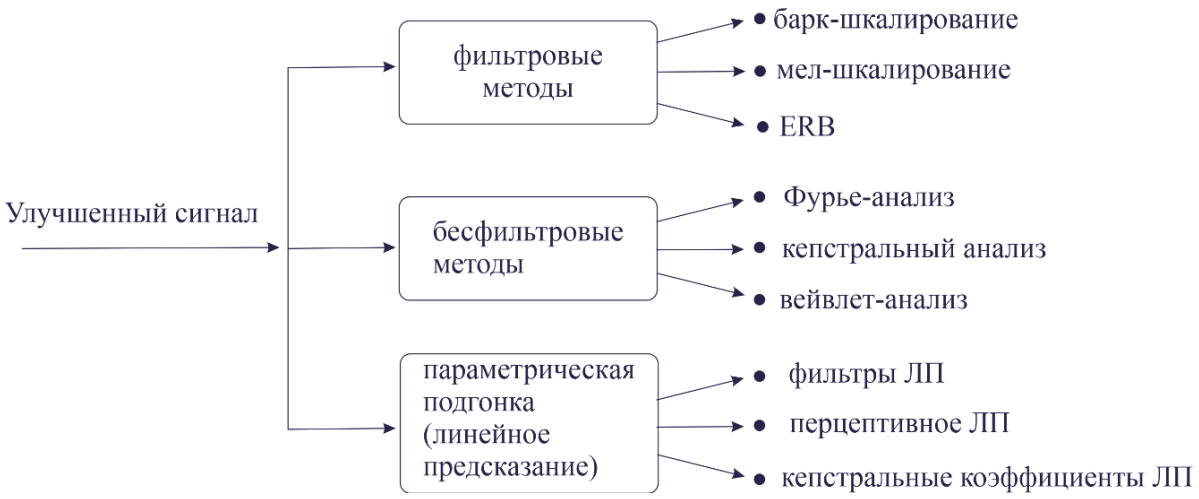

Рис. 1. *Основные методы и алгоритмы спектрального анализа*

Методы полосового анализа начали разрабатываться для применения еще в аналоговых схемах, но используются до сих пор, с наибольшей эффективностью – в системах компрессии звуковых сигналов и слухового протезирования. Методы линейного предсказания были разработаны в 70-х годах и оставались доминирующей техникой до ранних 80-х. В настоящее время и преобразование Фурье, и коэффициенты линейного предсказания широко применяются в различных речевых приложениях. Вейвлетное преобразование в настоящее время находится в стадии активной разработки.

2.1. Методы шкалирования полос

Теоретической основой для создания аналитической техники, имитирующей начальные этапы обработки речи периферической слуховой системой человека, послужила комбинация двух тезисов. Во-первых, базовой гипотезы слуховой «теории места» о пропорциональной связи максимума смещения базилярной мембраны и логарифма чистых тонов стимула [5]. Во-вторых, экспериментальных исследований частотной избирательности человеческого слуха методом режекторного шума, который сводится к обнаружению порога синусоидального сигнала, центрированного в спектральном пике шума, как функции ширины узкополосного шума. Последний показал, что частоты сложного звука в пределах определенной полосы пропускания некоторой номинальной частоты не могут быть индивидуально идентифицированы [6]. В случае же, когда одна из компонент сложного звука выпадает из полосы пропускания, она может быть идентифицирована.

Для описания частотной селективности человеческого слуха были предложены три шкалы, которые дают возможность выявить характеристики, идентифицирующие речь.

Э. Цвикером в 1957 году была предложена ставшая канонической концепция критических полос – деление частотного диапазона на критические полосы слуха – барки, а затем – набор стандартных значений для связи частоты и шкалы барков в табличной форме [7]. Предложен-

ная в [7] шкала барков делит частотный диапазон 20 Гц ~ 15.5 кГц на 24 критические полосы, границы которых определены экспериментально и представлены в табличной форме:

$$f \in \{100, 200, \dots, 12000, 1550\}, z \in \{1, 2, \dots, 23, 24\}.$$

Поскольку на практике использование таблиц не всегда удобно, Э. Цвикер, а затем другие исследователи предложили аналитические уравнения для перехода между частотами и шкалой барков, аппроксимирующие табулированные значения, а также для вычисления ширины критической полосы [7]:

$$Bark_{Zwicker}(f) = 13tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5tan^{-1}\left(\left(\frac{f}{7500}\right)^2\right). \quad (1)$$

Для данного выражения максимальное значение абсолютного отклонения от исходных табличных данных не превышает 0.2 барка в частотном диапазоне 20 Гц ~ 15.5 кГц.

Ширина критической полосы может быть оценена следующей формулой:

$$B_{Zwicker} = \frac{52548}{f_{bark}^2 - 52.56 + 690.39}. \quad (2)$$

В работе [8] предлагается аппроксимация частотного диапазона с помощью функции гиперболического синуса:

$$Bark_{Schroeder}(f) = 7sinh^{-1}\left(\frac{f}{650}\right). \quad (3)$$

Для данного выражения точность аппроксимации составляет 0.2 барка в диапазоне 20 Гц ~ 4.2 кГц и снижается далее в области высоких частот.

Х. Траунмюллер ввел упрощенный набор аналитических выражений для применения концепции критической пропускной способности в речевых технологиях [9]:

$$Bark_{Traunmuller}(f) = 26.81(1 + \frac{1960}{f})^{-1} - 0.53. \quad (4)$$

Максимальное отклонение при использовании данного выражения составляет 0.05 барка в диапазоне 200 Гц ~ 6.8 кГц, а в диапазоне 100 Гц~ 8.2 кГц – 0.2 барка.

В работе [10] предложена формула для перехода к шкале барков, обладающая высокой точностью в широком диапазоне 20 Гц ~ 15.5 кГц:

$$Z_{Kaval'chuk}(f) = a \cdot ln\left(c_1 + m \cdot ln\left(c_2 + \left(\frac{f+o}{p}\right)^q\right)\right), \quad (5)$$

где $a = 8.96$, $c_1 = 0.978$, $m = 5$, $o = 75.4$, $p = 2173$, $q = 1.347$, $c_2 = 0.994$.

Критерием удовлетворительности той или иной аппроксимации является не только оценка абсолютного отклонения, но также простота аналитического выражения и его обратимость. Кроме того, аппроксимирующие выражения оцениваются в соответствии с целями применения в разного рода речевых приложениях и психоакустических исследованиях. Наиболее широко применяемыми являются наборы аналитических выражений, предложенные Цвикером, Шредером и Траунмюллером, – формулы (1)–(4). Несмотря на максимальную абсолютную погрешность в 0.2 барка, для анализа и обработки звуковых сигналов широкого диапазона чаще всего применяется выражение (1), предложенное Цвикером. Для обработки узкополосных и речевых сигналов чаще используются выражения (3) и (4).

Шкала мелов представляет собой альтернативное отображение акустической частоты f на шкалу перцептивно значимых частот и определяется следующим образом [3, 11]:

$$mel = 2595 \cdot lg\left(1 + \frac{f}{700}\right). \quad (6)$$

Данное преобразование разработано с учетом нелинейных особенностей восприятия высоты тона [3, 11] и чаще используется для обработки музыкальных произведений. Мел аппроксимируется как линейная шкала в диапазоне от 0 до 1000 Гц, а затем в более высоком диапазоне частот как логарифмическая.

Наконец, частотная селективность человеческого слуха может быть описана в терминах эквивалентно-прямоугольных полос пропускания (*ERB*) на основе концепции, предложенной Муром и Гласбергом [12, 13]. Эквивалентная прямоугольная полоса пропускания представляет собой упрощенную аппроксимацию человеческого слуха идеальными прямоугольными полосовыми фильтрами на основе экспериментальных исследований с одновременным маскированием без учета суммирования громкости.

Сначала Мур и Гласберг ввели полиномиальную аппроксимацию ширины слуховых фильтров человека для звуковых стимулов умеренной громкости и молодых слушателей, а затем предложили линейную [12]:

$$ERB(f) = 24.7(4.37f + 1), \quad (7)$$

применимую для звуков умеренной громкости в диапазоне 0.1~10 кГц.

Шкала пропускной способности может быть определена как функция *ERBS(f)*, возвращающая число полос пропускания ниже заданной частоты. При использовании линейной аппроксимации функция имеет следующий вид [13]:

$$ERBS(f) = 21.4log_{10}(1 + 0.00437f). \quad (8)$$

Модельные представления *ERB* не идентичны классической концепции критических полос, хотя и связаны с ней количественно. *ERB* можно считать мерой частотного разрешения, а барк-шкалу – мерой тонотопической позиции. Применение *ERB* позволяет объяснить особенности восприятия, в том числе псевдо-логарифмический рост ширины критической полосы с ростом частоты и логарифмический закон восприятия интервалов частот. Несомненным достоинством *ERB* является невосприимчивость к биениям и интермодуляциям между сигналом и маскером [12].

Концепция критических полос широко используется, например, в слуховом преобразовании Фурье, кодировании речи, анализе сигналов и обработке сигналов для слухового протезирования (слуховые аппараты и кохлеарные имплантаты).

Очевидно, наиболее удобным инструментом для реализации описанных модельных представлений является полосовая фильтрация, которая дает возможность осуществить декомпозицию сигнала, понизить частоту дискретизации и строить эффективные системы частотной селекции сигналов. Таким образом, функционирование улитки можно рассматривать как банк фильтров, выходы которого упорядочены тонотопически вдоль барк- или мел-шкалы. Ширина полосы пропускания выбирается равной критической ширине, соответствующей центральной частоте.

Один из таких банков, предложенный в [7], стал неким стандартом в обработке речи. Центральные частоты и ширина полос для данных фильтров соответствуют тем частотам, для которых (1) принимает значение целого индекса в таблице Цвикера, ширина полосы пропускания вычисляется с помощью формулы, предложенной в [7].

Другой, равно важной фильтрацией в литературе, посвященной обработке речи, признана фильтрация на основе мел-шкалы [3]. В данном подходе линейно определены десять фильтров от 100 до 1000 Гц. Далее в области свыше 1000 Гц определены пять фильтров для каждого удвоения частотной шкалы. Указанные фильтры расположены логарифмически. Каждый фильтр в цифровом банке фильтров обычно реализуется как линейный фазовый фильтр таким образом, что групповая задержка всех фильтров равна нулю и исходящие сигналы от фильтров будут синхронизированы по времени. Уравнения фильтров для реализации линейного фазового фильтра могут быть записаны в следующем виде:

$$s_i(n) = \sum_{j=-\frac{N_{FBi}-1}{2}}^{\frac{N_{FBi}-1}{2}} \alpha_{FBi}(j)s(n+j), \quad (9)$$

где $\alpha_{FBi}(j)$обозначает j-й коэффициент для i-го критического полосового фильтра.

Порядок фильтра обычно нечетный для линейного фазового фильтра.

Выходные данные фильтрации обычно обрабатываются с помощью одного из методов оценки энергии [3], а затем комбинируются с другими параметрами для формирования вектора измерений сигнала.

Фильтрация, сначала аналоговая, затем цифровая, будучи самым ранним из подходов, созданных для обработки сигналов в распознавании речи, наиболее часто используется в системах, имитирующих слуховую обработку человеком. Это объясняется в первую очередь высокой эффективностью сжатия, осуществляемого на основе психоакустических исследований, в отличие от компрессии, реализуемой на основе анализа статистических свойств сигнала. Важнейшим преимуществом банков фильтров для применения в речевых технологиях является тот факт, что мощность выходов варьируется в зависимости от типа произносимого звука, то есть определенные выходные данные фильтров могут коррелировать с определенными классами речевых звуков. Кроме того, данный анализ основан на линейной обработке, что делает его устойчивым к шуму окружающей среды. Хотя в настоящее время фильтрация используется главным образом для компрессии сигнала, проектирование оптимальных банков фильтров способствует решению целого ряда прикладных задач. Системы сжатия разрабатываются для передачи звука высокого качества в сети Интернет. Значительная компрессия звука без снижения качества звучания необходима для расширения возможностей выбора частотных диапазонов в системах вещания и связи. Эффективная фильтрация позволит усовершенствовать шумоподавители, регуляторы уровня и тембра, усилители мощности и оптимизировать комплектацию музыкальных студий. Все еще актуальна задача понижения шума квантования, возникающего как следствие аналого-цифрового преобразования.

Современная методология проектирования банков фильтров не ограничивается методами расчета коэффициентов фильтров. Быстрыми темпами развиваются теория оптимального субполосного кодирования и теория построения банков фильтров основных компонент (Principal Component Filter Banks).

2.2. Методы бесфильтрового анализа

Наиболее применяемым методом акустического анализа данных при обработке речевых сигналов является преобразование Фурье [4, 13], которое декомпозирует сложный волновой процесс на элементарные гармонические колебания и описывает его свойства на основе тригонометрических функций.

Дискретное преобразование Фурье (ДПФ) оцифрованного сигнала вычисляется следующим образом [13]:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot \exp(-j\frac{2\pi kn}{N}),\ k = 0, \dots, N-1, \quad (10)$$

где $x[n]$ – цифровой аудиосигнал,

N – число отсчетов,

$\omega_k = \frac{2\pi k}{N}$ – угловая частота, рад.,

$f_k = f_s \cdot \frac{k}{N}$ – линейная частота, Гц,

f_s – частота дискретизации,

$X[k]$ – частотный спектр входящего сигнала,

k – частотный индекс.

При численной реализации ДПФ чаще всего используется алгоритм быстрого преобразования Фурье (БПФ) [4, 13]. Алгоритм БПФ представляет собой более эффективный способ вычисления ДПФ. Для большинства приложений длина преобразования выбирается равной степени двойки, что дает возможность выполнить в процессе расчетов $NlogN$комплексных сложений и $Nlog(\frac{N}{2})$комплексных умножений (для сравнения: в ходе вычисления ДПФ необходимо выполнить N^2 операций). Оборотной стороной преимущества по времени, обеспечиваемого БПФ, является необходимость настраивать нелинейные частотные отображения в соответствии с ортогональными частотными ограничениями БПФ.

Несомненно, дискретное преобразование Фурье является удобным средством анализа вследствие наглядности интерпретации его результатов, но предоставляет лишь обобщенные сведения о спектральном составе сигнала. Анализ сингулярностей невозможен, так как вычисление спектральных коэффициентов производится по частотному диапазону спектра в целом. Гармонические функции отображают специфические особенности сигнала (перепады бесконечной крутизны, разрывы, ступеньки, пики) только в отсутствии ограничений по числу членов ряда. На практике вследствие ограничения по числу членов ряда Фурье в окрестностях скачков и разрывов восстановленного сигнала возникают осцилляции.

Для решения указанных проблем используется кратковременное преобразование Фурье [13, 14]:

$$X_l[k] = \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} w[n] \cdot x[n + lH] \cdot \exp(-j\frac{2\pi kn}{N}),\ l = 0, 1, \dots, \quad (11)$$

где $w[n]$ – аналитическое окно,

x – фрагмент входящего сигнала,

l – индекс, указывающий порядковый номер фрейма (или временной индекс),

H – шаг.

Как показывает (11), кратковременное преобразование Фурье позволяет осуществить переход от амплитудного к частотно-временному представлению сигнала на основе свертки исследуемого сигнала и оконной функции. Вычисление БПФ на каждом интервале, выделяемом скользящей функцией аналитического окна, вскрывает особенности нестационарных сигналов. На современном этапе исследований в качестве аналитических оконных функций используются простейшие прямоугольные и треугольные окна, а также взвешивающие окна Хэмминга, Бартлетта, Гаусса, Кайзера, Ганна, Блэкмана, Кайера и др. [13], которые снижают искажения за счет выбора граничных условий. При обработке речи наиболее часто используемым является окно Хэмминга, которое обеспечивает высокое частотное разрешение и уменьшает уровень размытия спектра.

Временная разрешающая способность оконного преобразования определяется размером аналитического окна. Поскольку частотный спектр сигнала и его продолжительность связаны обратно пропорционально, то для выбранного значения частотного разрешения Δw ширина оконной функции вычисляется следующим образом: $\frac{2\pi}{\Delta w}$. На практике чаще всего размер окна определяется усредненным периодом основного тона (около 20 мс). В пределах окна сигнал считается стационарным.

Существуют алгоритмы, альтернативные быстрому преобразованию Фурье, которые служат тем же или подобным целям и могут иметь преимущества в некоторых случаях. Например, быстрое преобразование косинусов, дискретное преобразование Хартли [13], теоретико-числовое преобразование [13]. Дискретное преобразование косинусов широко используется в кодировании образов. Преобразование Хартли, оптимизированное для обработки реальных сигналов, не показало преимуществ перед БПФ [13]. Теоретико-числовое преобразование имеет специальную применимость для высокоточных вычислений.

К основным недостаткам Фурье-анализа можно отнести отсутствие хорошей частотно-временной локализации и усложненность параметрического представления сигнала, требующая дополнительной обработки и сжатия. Ввиду этого активно разрабатываются методы обработки речи, базирующиеся на вейвлет-преобразовании, а также приобрел популярность анализ кепстра.

Для отделения сигнала возбуждения от сигнала речевого тракта прибегают к кепстральному анализу. Анализ рассмотренных методов показал, что большинство из них сосредотачивают усилия на извлечении частотной характеристики речевого тракта человека, отбрасывая при этом характеристики сигнала возбуждения. Это объяснено тем, что коэффициенты первой модели обеспечивают лучшую разделимость звуков.

Кепстральный анализ был предложен в работе [15]. Кепстр – это спектр логарифма спектра временной волны:

$$c[n] = F^{-1}\{log|F\{x(n)\}|\},$$

где F и F^{-1} – прямое и обратное ДПФ.

На предварительном этапе обработки сигнал фильтруется с целью усиления высокочастотных составляющих спектра: $x_p(t) = x(t) - ax(t-1)$, $a \in [0.95, 0.98]$. Затем сигнал фреймируется и сглаживается оконной функцией, как правило Хэмминга. На следующем этапе осуществляется свертка спектра, полученного с помощью преобразования Фурье, со спектром принятого набора фильтров. Полученная огибающая спектра логарифмируется. И, наконец, применяется дискретное косинусное преобразование для вычисления кепстральных коэффициентов.

Данное преобразование позволяет представить спектр в сжатой форме.

Анализ кепстра послужил основанием для разработки целого семейства алгоритмов, среди которых наибольшее распространение получил метод мел-частотных кепстральных коэффициентов (MFCC).

Метод кепстральных мел-частотных коэффициентов (MFCC) первоначально был предложен для идентификации односложных слов в непрерывной речи, но не для идентификации говорящего. Вычисление MFCC является имитацией функционирования слуховой системы человека и предназначено для искусственного воплощения принципов работы уха. Принципы MFCC основаны на известной неравномерности расположения частотных фильтров человеческого уха, упорядоченных линейно на низких частотах и логарифмически на высоких частотах для сохранения фонетически важных свойств речевого сигнала. MFCC основан на дезинтеграции сигнала с помощью набора фильтров. Реализация MFCC сводится к дискретному косинусному преобразованию (DCT) реального логарифма кратковременного энергетического спектра, отображаемого на шкале частот Mel.

MFCC – это кепстральные коэффициенты, полученные на основе искаженной частотной шкалы, основанной на слуховом восприятии человека. При вычислении MFCC первое, что нужно сделать, – это оконная обработка речевого сигнала для разделения речевого сигнала на кадры. Поскольку высокочастотные форманты обрабатывают меньшую амплитуду по сравнению с низкочастотными формантами, высокие частоты усиливаются для получения одинаковой амплитуды для всех формант. После оконного анализа применяется быстрое преобразование Фурье (БПФ) для нахождения спектра мощности каждого кадра. Затем обработка банка фильтров выполняется по спектру мощности с использованием мел-шкалы. DCT применяется к речевому сигналу после преобразования спектра мощности в логарифмическую область для вычисления коэффициентов MFCC [5].

MFCC вычисляются с помощью следующего уравнения [16]:

$$\widehat{c_n} = \sum_{n=1}^{k} (log \widehat{s_k}) \cos\left(n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right), \quad (12)$$

где k – число мел-частотных коэффициентов,

$\hat{S}_k$ – выходные данные набора фильтров,

$\hat{C}$ – итоговые мел-частотные коэффициенты.

MFCC наиболее успешно применяется при создании внешнего интерфейса для приложений идентификации говорящего, поскольку позволяет повысить устойчивость к шумовым помехам с незначительной несогласованностью сеанса и легкостью для майнинга [16]. Кроме того, это MFCC безупречен при представлении звуков в тех случаях, когда характеристики источника стабильны и согласованы (музыкальные произведения и речь). Этот алгоритм извлекает информацию из дискретизированных сигналов с частотами максимум 5 кГц, что включает в себя большую часть энергии звуков, генерируемых людьми. Однако параметры, полученные на основе MFCC, не точны в условиях зашумленности [2] и могут не подходить для обобщения.

Потери релевантной информации в области высоких частот при вычислении кепстральных коэффициентов по мел-шкале было предложено компенсировать методом обратных мел-частотных кепстральных коэффициентов. Для их вычисления также используется набор треугольных фильтров на мел-шкале.

Одной из модификаций основного алгоритма MFCC является метод линейно-частотных кепстральных коэффициентов (LFCC), также использующий набор треугольных фильтров, но расположенных равномерно по линейной полосе частот.

Для более эффективного распознавания в условиях высокого зашумления был предложен метод кепстральных коэффициентов прямоугольного набора фильтров (RFCC), который осуществляет обработку критических полос слуха с помощью прямоугольных фильтров, распределенных по линейной частотной шкале без перекрытий.

Своего рода стандартом при моделировании частотной фильтрации человеческого слуха стало применение гамматон-фильтрации. Алгоритм гамматон-частотные кепстральные коэффициенты (GFCC – gammatone frequency cepstral coefficients) вычисляет кепстральные коэффициенты [17], используя импульсные характеристики гамматон-фильтра с центральной частотой f:

$$g(f,t) = \begin{cases} t^{a-1} e^{-2\pi f t} \cos(2\pi f t), t \geq 0, \\ 0 \end{cases} \quad (13)$$

где t – время,

a – порядок фильтра,

b – прямоугольная ширина полосы частот, которая возрастает с увеличением центральной частоты f.

Строго говоря, гамматон-фильтрация не является кепстральной обработкой, поскольку включает в себя операцию логарифмирования, функциональное сходство позволяет отнести данный метод к классу кепстральных.

Широкая практика применения кепстрального анализа продемонстрировала возможность решения задач распознавания речевых образов и идентификации дикторов с высокой эффективностью. К сожалению, кепстральный анализ частотных компонент свыше 1 кГц недостаточно эффективен вследствие большого расстояния между фильтрами в высокочастотном диапазоне [16]. Кроме того, логарифмирование повышает чувствительность к определенным видам шумов и искажений, поскольку является нелинейной процедурой обработки сигнала. По этой причине приложения, предназначенные для эксплуатации в зашумленных условиях, чаще всего используют параметрические методы подгонки спектра или кепстральные коэффициенты, получаемые из спектральных оценок высокого разрешения.

Как уже упоминалось выше, вследствие недостаточной информативности результатов Фурье-анализа во временном пространстве в настоящее время активно разрабатываются методы вейвлетного представления сигналов. Вейвлет-преобразование рассматривает исследуемый сигнал как суперпозицию масштабируемых базисных функций, хорошо локализованных во временной и частотной областях, а затем осуществляет последующую оценку корреляции анализируемого сигнала на текущем участке и версии вейвлета в выбранном масштабе [18]. То есть вейвлетное преобразование является обобщением спектрального анализа. Преобразование Фурье, расщепляющее сигнал на элементарные гармонические функции, является его частным случаем. Базисные вейвлетные функции по локализации в частном и временном представлении занимают промежуточное положение между гармоническими и импульсными функциями и имеют следующий вид:

$$\psi\left(\frac{t-b}{a}\right), \tag{14}$$

где b – параметр сдвига,

a – коэффициент масштабирования.

Вейвлетный анализ начинается с размещения базисной функции с коэффициентом масштабирования, равным единице (что соответствует наиболее сжатому вейвлету), в начале сигнала ($t=0$). Затем осуществляется свертка с исследуемым сигналом, интегрирование по пределам задания вейвлета и нормировка на $1/\sqrt{a}$. Полученный коэффициент корреляции характеризует точку *C(1, 0)* масштабно-временной плоскости преобразования. Затем вейвлет сдвигается вправо на величину, равную параметру сдвига b, и вычисляется коэффициент *C(1, b)*. Процедура повторяется до конца сигнала. Для перехода к вычислению следующей масштабной строки коэффициент масштабирования увеличивается и т. д. Очевидно, что коэффициент масштабирования можно интерпретировать как ширину аналитического окна кратковременного преобразования Фурье. Аналогично координатная ось b соответствует временной оси сигнала. Минимальный размер окна вейвлетного анализа не превышает периода наивысшей гармоники. Вектор коэффициентов корреляции, получаемый в результате вейвлет-преобразования, иллюстрирует сходство характера изменений исследуемого сигнала в текущий момент времени и вейвлета в текущем масштабе.

Выбор базисной функции, близкой к виду анализируемого сигнала, является нетривиальной задачей. При решении данной задачи довольно часто используется преобразование Гильберта-Хуанга. Обработка с использованием преобразования Гильберта-Хуанга осуществляется на двух уровнях. На первом уровне осуществляется декомпозиция сигнала на эмпирические моды [19]:

$$s(t) = \sum_{i=1}^{I-1} imf_i(t) + r_I(t), \quad (15)$$

$imf_i(t)$ – эмпирические моды,
r_I – остаток разложения, $i=1, 2,\ldots$,
I – номер эмпирической моды.
На втором – формирование спектра Гильберта на основе эмпирических мод [19]:

$$HHT(t) = \sum_{i=1}^{T} a_i^2\,(t) \exp(q \int \omega_k\,(t)dt), \quad (16)$$

$a_i(t) = \sqrt{imf_i(t)^2 + IMF_i(t)^2}$ – модуль мгновенного значения амплитуды эмпирической моды,
$imf_i(t)$ – эмпирические моды,
$IMF_i(t) = \frac{1}{\pi} \int \frac{imf_i(\tau)}{t-\tau} d\tau$ – сопряженный по Гильберту сигнал эмпирической моды,
τ – временной сдвиг, пропорциональный фазе сигнала,
$\omega(t) = 2\pi fj$ – циклические частоты каждой эмпирической моды,
j – мнимая единица.
Значения $a(t)$ и $w(t)$ для каждой эмпирической моды вычисляются из соотношения:

$$Z_i(t) = imf_i(t) + jIMF_i(t). \quad (17)$$

Представление Гильберта-Хуанга отображает сигнал в частотно-энергетически-временную область и вскрывает таким образом скрытые модуляции и области концентрации энергии.

При обработке речи наиболее часто используемыми являются вейвлеты Хаара, Добеши, «Мексиканская шляпа», комплексный базис [18]. В последние годы особенно популярным при решении задачи извлечения признаков стал метод константного q-преобразования (*CQT*) [20].

CQT-метод можно рассматривать как набор фильтров, неравномерно распределенных в частотном пространстве, и определить в следующем виде:

$$X^{CQ}(k,n) = \sum_{j=n-\frac{N[k]}{2}}^{n+\frac{N[k]}{2}} x(j) a_k^* \left(j - n + \frac{N[k]}{2}\right), \quad (18)$$

где k – индекс частотной полосы,
$N[k]$ – переменный размер окна,
a^*_k – комплексно сопряженные базисные функции.
Базисные вейвлеты в данном случае имеют вид:

$$a_k[n] = \frac{1}{C}\left(\frac{n}{N[k]}\right) \exp\left[i\left(2\pi n \frac{f[k]}{f_s}\right) + \Phi[k]\right], \quad (19)$$

$f[k]$ – критическая частота k-й полосы,
f_s – частота дискретизации,
$w(t)$ – оконная аналитическая функция,
$\Phi[k]$ – сдвиг фазы,
C – коэффициент масштабирования,
$N[k] = \frac{f[k]}{f_s} Q$,
$Q = \frac{f[k]}{f[k-1]-f[k]}$.

Таким образом, вейвлет-преобразование предоставляет более информативный анализ нестационарных во временном и неоднородных в частотном представлениях сигналов по сравнению с остальными методами спектрального анализа, в том числе по сравнению с широко применяемым преобразованием Фурье. Вейвлет-спектрограммы, представляющие собой двумерные поверхности в масштабно-временном пространстве, позволяют анализировать как глобальные, так и локальные свойства сигналов. Масштабирование позволяет представить сигнал с разной степенью детализации и выделить в разномасштабных процессах лишь те уровни, которые представляют интерес. Сдвиг дает возможность с высокой точностью анализировать локальные особенности сигналов (скачки, провалы, разрывы, ступеньки или всплески). Кроме того, вейвлетное представление в отличие от преобразования Фурье, использующего для описания сигнала лишь одну функцию $e^{j\omega t}$, обладает достаточно широким разнообразием функций, свойства которых можно гибко ориентировать на решение различных задач. Однако говорить о том, что в ближайшем будущем вейвлет-анализ заменит преобразование Фурье, еще рано, поскольку критерии разработки оптимальных алгоритмов еще не достигли необходимого уровня обобщения. При практической реализации вейвлет-методов исследователям приходится уделять большое внимание корректности и эффективности их работы.

2.3. Методы линейного предсказания

Проанализируем теперь линейное предсказание – широкий класс методов параметрической подгонки спектра исследуемого сигнала для вычисления его фундаментальных составляющих. В отличие от моделей, основанных на линейном спектральном анализе, параметрическое моделирование рассматривает построение модели спектра как авторегрессивный процесс. Линейное предсказание основано на акустической модели речепроизводства Фанта [21], которая рассматривает речевой сигнал как выход линейной системы с переменными параметрами, возбуждаемой либо квазипериодическими импульсами, либо стохастическим шумом. Математически речевой сигнал как свертка функции возбуждения и набора линейных фильтров описывается следующим уравнением [22]:

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + Gu(n), \tag{20}$$

где $\{a_k\}$ – коэффициенты цифрового фильтра,

G – коэффициент усиления,

$u(n)$ – сигнал возбуждения.

Линейное предсказание отыскивает такую модель сигнала, на выходе которой каждый текущий отсчет является линейной комбинацией предыдущих отсчетов:

$$\tilde{x}(n) = \sum_{k=1}^{\mathrm{p}} \alpha_k x(n-k) + \varepsilon(n), \tag{21}$$

где p – число коэффициентов модели предсказателя,

$\{\alpha_k\}$ – коэффициенты линейного предсказания,

$\varepsilon(n)$ – погрешность предсказания.

$$\varepsilon(n) = x(n) - \tilde{x}(n) = x(n) - \sum_{k=1}^{p} \alpha_k x(n-k). \tag{22}$$

Если искомый сигнал удовлетворяет модели, а также $a_k = \alpha_k$, сравнивая (20) и (21), получим

$$\varepsilon(n) = Gu(n). \tag{23}$$

То есть, как показывает (23), фильтр погрешности предсказания является обратным фильтром линейной системы. Поэтому, минимизируя среднеквадратичную погрешность предсказания, можно с высокой точностью оценить параметры линейной системы. На практике задача линейного предсказания сводится к вычислению набора коэффициентов линейного предсказания. Существует три основных способа вычислить коэффициенты предсказания: ковариантные методы на основе ковариантной матрицы (или метод наименьших квадратов) [22], методы автокорреляции [22] и лестничных фильтров (гармонические) [22]. В распознавании речи почти эксклюзивно используется метод автокорреляции вследствие эффективности и стабильности.

Концептуальные представления анализа линейного предсказания послужили основой для создания трех модификаций данной модели, на основе которых, в свою очередь, были разработаны целые наборы высокоэффективных методов извлечения признаков. Это кепстральные коэффициенты линейного предсказания, перцептивное линейное предсказание и фильтры линейного предсказания.

Фильтрация на основе метода коэффициентов линейного предсказания, по существу, представляет собой комбинацию базового понятия критической полосы банка фильтров и модели линейного предсказания. Данный подход предоставляет для последующей обработки и классификации амплитуды банка фильтров, полученные на основе фильтрации выборки из модели линейного предсказания (а не спектра сигнала) на соответствующих частотах банка фильтров. Простейший способ вычисления амплитуд банка фильтров включает в себя прямую оценку модели линейного предсказания [3]:

$$S_{LP}(f) = \frac{G_{LP}}{\sum_{i=0}^{N_{LP}} \alpha_{LP}(i)\exp(-j2\pi\frac{f}{f_S})}, \tag{24}$$

где f_s – частотная выборка.

Данный метод требует около $(4p + 3)$операций сложения/умножения на одну частотную выборку. Обычно спектр передискретизируется, поэтому для вычисления амплитуд банка фильтров генерируются усредненные оценки.

Другим популярным подходом является вычисление энергетического спектра из автокорреляции импульсного ответа, вычисляемого напрямую из коэффициентов линейного предсказания [22]:

$$R_{LP} = \begin{cases} \sum_{m=0}^{N_{LP}-|k|} \alpha_{LP}(m)\alpha_{LP}(m + |k|), |k| \le N_{LP}. \\ 0, \ |k| > N_{LP} \end{cases} \tag{25}$$

Спектральная плотность мощности может быть вычислена эффективно из функции автокорреляции с учетом того, что функция автокорреляции четная действительная функция. Значит, ее преобразование Фурье действительно и определяется следующим образом:

$$S_{LP} = \begin{cases} R_{LP}(0) + 2\sum_{k=1}^{N_{LP}} R_{LP}(k)\cos(2\pi\frac{f}{f_s}k), |k| \le N_{LP}. \\ 0, \ |k| > N_{LP} \end{cases} \tag{26}$$

Уравнение (25) требует в общей сложности $N_{LP}^2 - \frac{3}{2}N_{LP}$ операций сложения/умножения, а (26) – N_{LP} операций сложения/умножения на выборку. При любом подходе нелинейно искаженные спектры могут быть легко реализованы путем соответствующих выборов частот выборки банка фильтров. Также хотя модель линейного предсказания подразумевает подгонку сглаженного спектра, передискретизация спектра оказывается удобной для точной характеристики резких пиков в частотном ответе банком фильтров (который имеет тенденцию грубо квантовать спектр).

Следующая модификация модельных представлений линейного предсказания сводится к использованию линейного предсказания для вычисления кепстральных коэффициентов, реализуемому с помощью рекурсивной функции [3]:

$$c(n) = \begin{cases} 0, n < 0, \\ \ln(A), n = 0, \\ a_n - \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k}, 0 < n < p, \\ \sum_{k=n-p}^{n-1} \frac{k}{n} c(k)\, a_{n-k}, n > p. \end{cases} \quad (27)$$

Существует небольшое затруднение в вычислении рекурсивных кепстральных коэффициентов. Поскольку кепстральные коэффициенты фактически являются обратным преобразованием Фурье импульсного ответа линейного предсказания, а ЛП в свою очередь – фильтр бесконечного импульсного ответа, существует теоретическая возможность вычислить бесконечное число кепстральных коэффициентов. При практической реализации чаще всего число кепстральных коэффициентов сравнимо с числом коэффициентов линейного предсказания.

Кепстральные коэффициенты линейного предсказания обладают сниженной по сравнению с коэффициентами линейного предсказания частотой ошибок. Широкое практическое использование кепстральных оценок продемонстрировало их устойчивость к шумовым загрязнениям [3]. Нужно отметить, что это важное преимущество, поскольку линейное предсказание, будучи нелинейной обработкой сигнала, указанным преимуществом не обладает. Точнее, кепстральные коэффициенты имеют разную чувствительность к разным видам зашумления: коэффициенты более низкого порядка чувствительны к наклону спектра, а коэффициенты более высокого порядка чувствительны к шуму.

Метод перцептивного линейного предсказания разрабатывался для исключения индивидуальных особенностей дикторов при решении задачи распознавания речи и оценки слуховой информации по значимости с учетом результатов психолингвистических исследований, а именно: критических полос слуха с маскированием, кривой равной громкости, степенной связи между громкостью и интенсивностью звука. Перцептивное линейное предсказание, воспроизводя характерные особенности слуховой обработки человека, предоставляет сжатый сглаженный кратковременный спектр речи, очень сходный с тем, который периферическая слуховая система человека передает для обработки в подкорковые зоны и высшие отделы мозга.

Для вычисления коэффициентов перцептивного линейного предсказания каждый речевой сигнал пофреймово анализируется с помощью кратковременного преобразования Фурье (с использованием окна Хэмминга и БПФ), затем вычисляется сумма квадратов действительной и мнимой Фурье-компонент, то есть определяются спектральные оценки мощности. Частоты спектра мощности ω пересчитываются в барк-шкалу:

$$\Omega(\omega) = 6ln\left[\frac{\omega}{1200\pi} + \left(\frac{\omega^2}{1200\pi} + 1\right)^{0.5}\right]. \quad (28)$$

Затем с интервалом в 1 барк применяется трапециевидный фильтр для сжатия высоких частот вследствие интеграции перекрывающихся характеристик фильтра критической полосы в спектре мощности. Последующая симметричная свертка частотной области по искаженной барк-шкале позволяет имитировать маскировку высоких частот низкими, одновременно сглаживая спектр. Далее спектр сглаживается функцией кривой равной громко-

сти для имитации неравномерной чувствительности человеческого слуха на различных частотах на уровне 40 дБ:

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)]. \quad (29)$$

Наконец, по закону Стивенса оценивается громкость сигнала путем извлечения кубического корня из амплитуды спектра.

На заключительном этапе выполняется обратное дискретное преобразование Фурье для получения коэффициентов автокорреляции, спектральное сглаживание на основе решения уравнения авторегрессии. Коэффициенты авторегрессии конвертируются в кепстральные переменные.

PLP был разработан с целью избавления от информации, связанной с индивидуальными особенностями дикторов, но показал более эффективную классификацию по сравнению с линейным предсказанием не только в данном аспекте, но и высокую устойчивость к вариативности, обусловленной качеством микрофонов и каналами передачи [52, 53]. Кроме того, PLP-анализ обладает довольно низкой чувствительностью к спектральному наклону, что согласуется с выводами о том, что он относительно нечувствителен к фонетическим оценкам спектрального наклона [7].

Завершая анализ линейного предсказания в целом, необходимо отметить важное значение параметрического моделирования в области обработки речи с момента его разработки в начале 70-х [3]. Благодаря математически точной и простой в реализации модели параметры линейного предсказания практически сразу же стали использоваться при решении задачи распознавания в речевых системах. К концу 70-х почти каждая система обработки речи использовала тот или иной алгоритм параметрической подгонки спектра для приложений, нацеленных на распознавание, сжатие или верификацию. Применение LPC обеспечивает качественное разделение источника и вокального тракта и как следствие – простое представление его характеристик. Поскольку анализ линейного предсказания – нелинейная операция, распознавание в зашумленных условиях усложняется. По этой причине некоторые системы все еще используют анализ банка фильтров, основанный на преобразовании Фурье. Особенно эффективное распознавание модель демонстрирует на озвученных участках речи [22]. Использование линейного предсказания на переходных участках менее эффективно.

Основным затруднением реализации данных модельных представлений является необходимость вычисления коэффициентов в пределах короткого временного интервала. Отметим, что с уменьшением продолжительности спектра (и окна) временное разрешение в спектрограмме возрастает. Чаще всего в распознавании речи используется продолжительность фрейма 20 мс. В последнее время, поскольку фокус исследований сместился в сторону фонетического распознавания, продолжительность фрейма в 10 мс стала крайне распространенной. Смещение в сторону повышения разрешения по времени продолжится с развитием речевых технологий.

В настоящее время параметрические модели чаще используются для компрессии и реконструкции речи. Метод LPC применяется также для создания мобильных роботов, при осуществлении тонального анализа скрипок и других струнных музыкальных инструментов. Кроме того, на основе представлений линейного предсказания было создано целое семейство методов извлечения признаков – это кепстральные коэффициенты линейного предсказания (LPCC), логарифмическое соотношение площадей (LAR), коэффициенты отражения (RC), линейные спектральные частоты (LSF) и синусоидальные коэффициенты Arcus (ARCSIN).

Все исследованные выше методы параметризации речевых сигналов экстрагируют признаки в пределах кратковременного фрейма. Для дополнения абсолютных признаков при-

знаками, характеризующими динамику сигнала, векторы дополняют первыми и вторыми производными Δ- и Δ-Δ-коэффициентов.

Далее, на стадии постобработки осуществляются конкатенация, нормализация и декорреляция признаков. Среди методов нормализации наиболее широкое применение получил метод вычитания кепстрального среднего, снижающий негативное влияние каналов передачи [3].

3. Заключение

Разработка и применение методов извлечения признаков требуют большого внимания, так как эффективность распознавания в первую очередь определяется данной фазой обработки. Однако даже исчерпывающий обзор существующих техник анализа не укажет универсальный алгоритм, поскольку вектор, наиболее полно характеризующий звуковое сообщение, определяется задачей распознавания и выбором классификатора. Систематическое исследование современных методов извлечения акустических признаков речи, скорее, поможет сориентироваться в выборе технологии при проектировании столь важного элемента речевой системы. Обобщенные научные принципы выбора того или иного алгоритма извлечения признаков сводятся к необходимости сохранить те параметры, по которым элементы звукового сигнала могут быть дифференцированы с высокой надежностью; выявить перцептивно значимые аспекты речевого сообщения, устойчивые к вариативности разной природы и характеризующие динамику спектра. Можно отметить, что наиболее популярным методом цифровой обработки речевых сигналов как при разработке автоматических систем распознавания речи, так и при решении задачи распознавания диктора являются мел-частотные кепстральные коэффициенты [3]. В последние годы в качестве альтернативы MFCC активно разрабатываются искусственные нейронные сети с использованием так называемых bottleneck-признаков, позволяющие повысить устойчивость к вариативности относительно индивидуальности речевого тракта. Методы преобразования Фурье традиционно признаются надежными в сильно зашумленных условиях и популярны из-за сходства с начальными этапами обработки слуховой системой человека. Довольно часто при параметризации речи разработчики используют комбинацию из нескольких техник. Чаще всего такие комбинации применяют при решении задачи противодействия спуфинговым атакам.

Разработка методов цифровой обработки сигнала развивается в направлении использования статистических свойств высокого уровня, например, асимметрии и эксцесса вследствие того, что распределение речевого сигнала во временной и частотной областях является негауссовым.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. *Deller J.R., Proakis J.G., Hansen J. H.L.* Discrete Time Processing of Speech Signals. Hoboken NJ: Wiley-IEEE Press, 1999. 936 p.

2. *Gupta V.* A Survey of Natural Language Processing Techniques. International Journal of Computer Science & Engineering Technology (IJCSET). 2014. Vol. 5. No. 1. Pp. 14–16.

3. *Picone J.W.* Signal modeling techniques in speech recognition. Proceedings of the IEEE. 1993. Vol. 81. No. 9. Pp. 1215–1245.

4. *Jurafsky D., Martin J.* Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. NJ: Prentice Hall, 2009. 1024 p.

5. *Pickles J.O.* An Introduction to the Physiology of Hearing. New York: Academic Press, 1988. 400 p.

6. *Fletcher H., Munson W.A.* Relation between Loudness and Masking. J. Acoust. Soc. Am. 1937. No. 9. Pp. 1–10.

7. *Zviker E., Feldkeller R. Ukho kak priyemnik informatsii* [Ear as a receiver of information]. Moscow: Svyaz', 1971. 255 p. (In Russian)

8. *Schroeder M.R.* Optimizing Digital Speech Coders by Exploiting Masking properties of the Human Ear. J. Acoust. Soc. Am. 1979. No. 66(6). Pp. 1647–1652. https://doi.org/10.1121/1.383662

9. *Traunmuller H.:* Analytical Expressions for the tonotopic sensory scale. The Journal of the Acoustical Society of America. 1990. Vol. 88. No 1. Pp. 97–100. DOI: 10.1121/1.399849.

10. *Kavalchuk A.N.* The formula for the transition from the frequency domain to the bark scale and vice versa. *Informatika* [Informatics]. 2011. No. 4(32). Pp. 71–81. (In Russian)

11. *Aldoshina I., Pritts R. Muzykal'naya akustika* [Musical acoustics]. Textbook. St. Petersburg: Composer, 2014. P. 720. (In Russian)

12. *Moore B.* Frequency selectivity in hearing. Boston, MA: Springer, 1986. P. 456. https://doi.org/10.1007/978-1-4613-2247-4

13. *Smith J.O.* Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications. W3K Publishing. 2007. P. 322 http://books.w3k.org/

14. *Rabiner L.R., Schafer R.W.* Digital processing of speech signal. New Jersey: Prentice-Hall, 1978. P. 496 (Russ. ed.: Rabiner L.R., Shafer R.V. *Tsifrovaya obrabotka rechevykh signalov.* Moscow: Radio i svyaz' Publ., 1981. 496 p.).

15. *Davis S., Mermelstein P.* Experiments in syllable-based recognition of continuous speech. IEEE Transactions on Acoustics, Speech and Signal Processing. 1980. Vol. 28. Pp. 357–366.

16. *Chakroborty S., Roy A., Saha G.* Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification. IEEE International Conference on Industrial Technology. Mumbai, 2006. Pp. 387–390.

17. *Shao Y., Wang D.L.* Robust speaker identification using auditory features and computational auditory scene analysis. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008). Las Vegas, NV, USA, 2008. Pp. 1589–1592. DOI: 10.1109/ICASSP.2008.4517928.

18. *Novikov L.V. Osnovy veyvlet-analiza signalov* [Fundamentals of Wavelet Signal Analysis]. Tutorial. St. Petersburg: IanP RAN, 1999. 152 p. (In Russian)

19. *Huang N.E.* The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London A. 1998. Vol. 454. Pp. 903–995.

20. *Todisco M., Delgado H., Evans N.* A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients. Odyssey 2016. The Speaker and Language Recognition Workshop. Bilbao, Spain, 2016. Pp. 283–290.

21. *Fant G.* Acoustic Theory of Speech Production. Walter de Gruyter, 1970. P. 328.

22. *Rabiner L., Juang B.-H.* Fundamentals of speech recognition. NJ: Prentice-Hall, Inc., 1993. P. 507.

Информация об авторах

Гуртуева Ирина Асланбековна, науч. сотр. отдела «Компьютерная лингвистика», Институт информатики и проблем регионального управления – филиал Кабардино-Балкарского научного центра РАН;

360000, Россия, Нальчик, ул. И. Арманд, 37-а;

gurtueva-i@yandex.ru, ORCID: https://orcid.org/0000-0003-4945-5682

Бжихатлов Кантемир Чамалович, канд. физ.-мат. наук, зам. директора по науке, Институт информатики и проблем регионального управления – филиал Кабардино-Балкарского научного центра РАН;

360000, Россия, Нальчик, ул. И. Арманд, 37-а;

haosit13@mail.ru, ORCID: https://orcid.org/0000-0003-0924-0193

MSC: 68T10 Original article

ANALYTICAL REVIEW AND CLASSIFICATION OF METHODS FOR FEATURES EXTRACTION OF ACOUSTIC SIGNALS IN SPEECH SYSTEMS

I.A. GURTUEVA, K.Ch. BZHIKHATLOV

Institute of Computer Science and Problems of Regional Management –
branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences
360000, Russia, Nalchik, 37-a I. Armand street

Annotation. This paper presents an overview of methods and algorithms for feature extraction to transform an acoustic signal into a sequence of vectors for solving problems of segmentation, classification, identification, or speech recognition. A classification of feature extraction methods according to mathematical approaches is proposed. The algorithms and techniques of spectral analysis, which are most used in the design of speech recognition systems, are discussed. This review clearly demonstrates the complexity of the problem of acoustic processing - searching a representation that decreases the dimension of the model and maintain the completeness of linguistic information and, importantly, is stable to variability with respect to the speaker, transmission channels and the environment. The analysis of the existing feature extraction methods is useful for selection of a technology when designing a key element of a speech system.

Keywords: speech recognition, Fourier analysis, cepstral analysis, linear prediction, methods for feature extraction

The article was submitted 08.02.2022 *Accepted for publication 14.02.2022*

For citation. Gurtueva I.A., Bzhikhatlov K.Ch. Analytical review and classification of methods for features extraction of acoustic signals in speech systems. News of the Kabardino-Balkarian Scientific Center of RAS. 2022. No. 1 (105). Pp. 41–58. DOI: 10.35330/1991-6639-2022-1-105-41-58

Information about the authors

Gurtueva Irina Aslanbekovna, Researcher of the laboratory of Computer Linguistics, Institute of Compute Science and Regional Management Problems – branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences;

360000, Russia, Nalchik, 37-a I. Armand street;

gurtueva-i@yandex.ru, ORCID: https://orcid.org/0000-0003-4945-5682

Bzhikhatlov Kantemir Chamalovich, Candidate of Physics and Mathematics sciences, Deputy Science Director, Institute of Compute Science and Regional Management Problems – branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences;

360000, Russia, Nalchik, 37-a I. Armand street;

haosit13@mail.ru, ORCID: https://orcid.org/0000-0003-0924-0193