Удалённая система автоматического распознавания речи

Маковкин К.А.,
сотрудник Вычислительного центра им. А.А. Дородницына

В работе описаны принципы построения систем автоматического распознавания речи в среде цифровых компьютерных сетей. Рассмотрены проблемы, возникающие при реализации такой архитектуры системы распознавания, и приведены способы их решения.

This paper introduces the basic features of speech recognition over digital networks. The rapid development of various digital networks (mobile, wireless and IP-based) during the last few years has opened a new area of expansion for speech recognition technologies and has allowed developing of remote speech recognition systems. Remote speech recognition systems rely on efficient transmission of speech information from remote clients to a centralized server. There are two approaches of remote speech recognition systems realization: network speech recognition and distributed speech recognition. However, the relief of computational demand on remote devices comes at the cost of network deteriorations and additional components such as feature quantization, error recovery and concealment. All these issues have been recently addressed in a series of ETSI standards. In addition, this paper introduces VoIP protocols, which provides session establishment and management functions of speech recognition server.

Ключевые слова: автоматическое распознавание речи, удалённое распознавание речи, сетевое распознавание речи, распределённое распознавание речи, скрытые марковские модели, речевые кодеки, протоколы VoIP.

Введение

Стремительное развитие цифровых компьютерных сетей привело к значительном росту числа распределённых информационных систем, которые

позволяют пользователям обращаться к ним с различными интерактивными запросами (заказ билетов, получение справки о состоянии банковского счёта или получение другой справочной информации). Это сопровождается широким распространением мобильных устройств, которые используются как персональные цифровые помощники, — это карманные персональные компьютеры (КПК), мобильные телефоны и смартфоны, GPS-навигаторы, нетбуки и т. д. Так, по информации интернет-сайта InformaTM (http://www.informatm.com), число пользователей одних только мобильных телефонов в 2010 г. достигнет 3,5 миллиарда. При этом разработчики мобильных устройств стремятся снабдить практически все свои изделия широким спектром различных сетевых проводных (Ethernet) и беспроводных интерфейсов (GPRS, Wi-Fi, Bluetooth и т. д.), что позволяет пользователю такого устройства получить доступ к сетевым ресурсам из любого места и в любое время.

В то же время, несмотря на значительные достижения в области разработки человеко-машинных интерфейсов, речевое взаимодействие остаётся наиболее удобным и комфортным для человека. В условиях повсеместного использования компьютеров становится очевидно, что взаимодействие с устройством с помощью клавиатуры или стилуса, а также отображение информации на маленьком экране — ограниченный и крайне неудобный для человека интерфейс по сравнению с привычным речевым общением. Это приводит к росту интереса к разработке и внедрению систем автоматического распознавания речи (АРР), так как это повысит удобство использования системы или устройства. Примером может служить применение АРР в такой системе, как интерактивный речевой ответ (Interactive voice response, IVR), который широко используется в телефонных сетях. Возможность общаться с системой на естественном языке избавляет пользователя от неудобного и сложного взаимодействия с помощью тонального набора цифр на клавиатуре телефона, которое часто приводит к ошибочному выбору желаемого режима. В то же время использование речевого интерфейса, наоборот, позволяет пользователю легко перемещаться по меню любой сложности. В результате сокращается продолжительность телефонного звонка и повышается качество обслуживания клиента.

К сожалению, большинство существующих систем АРР, ориентированных на распознание слитной речи, с большими словарями без подстройки под диктора разрабатывались для работы на настольных рабочих станциях или для мощных компьютеров и поэтому плохо подходят для работы на нетбуке или даже на мобильном устройстве. Это связано, прежде всего, с тем, что при разработке мобильных устройств основные цели — снизить цену устройства, уменьшить его размер и увеличить время автономной работы. Поэтому разработчики стремятся минимизировать вычислительные ресурсы, т. е. использовать центральные процессоры с минимальным энергопотреблением (соответственно с пониженной тактовой частотой), относительно небольшой размер оперативной памяти с низкой скоростью обмена, а также небольшой размер постоянной энергонезависимой памяти. Очевидно, что такие ограничения на вычислительные ресурсы делают адаптацию систем АРР для работы на мобильном устройстве крайне сложной. В этой ситуации коммуникационные возможности мобильных устройств оказываются очень важными и наводят на мысль о разделении системы АРР на несколько задач, которые смогут выполняться на разных компьютерах. Другими словами, построить АРР на основе клиент-серверной архитектуры, где система АРР располагается на некотором мощном удалённом сервере, а клиентский компьютер осуществляет ввод и передачу оцифрованного речевого сигнала в виде запроса на распознавание по цифровым компьютерным сетям. Таким образом, пользователь, располагающий незначительными вычислительными ресурсами, получает возможность использовать полноценную систему АРР и взаимодействовать с устройством или информационной системой на естественном языке. По принципу построения такая система называется удалённой системой распознавания речи (УСРР) (**рис. 1**).

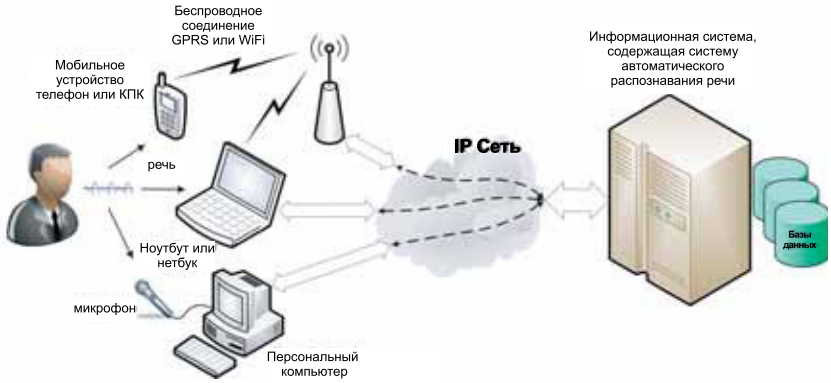

Рис. 1. Использование удалённой системы распознавания речи

Система автоматического распознавания речи

На сегодняшний день наиболее распространённые и успешные системы АРР, построенные на принципах скрытых марковских моделей (СММ). В рамках такого подхода речевой сигнал представляется как случайный образ, который необходимо распознать или, другими словами, преобразовать в некоторую последовательность слов W. Тогда задача распознавания речевого сигнала формулируется как классическая задача классификации образов по критерию максимума апостериорной вероятности, т.е. необходимо максимизировать апостериорную вероятность $P(W|X)$, где X — наблюдаемая последовательность акустических векторов параметров речевого сигнала, а W — последовательность слов. Согласно формуле Байеса апостериорную вероятность можно переписать в виде

$$\arg\max_{W\in\Gamma} P(W \mid X) = \arg\max_{W\in\Gamma} P(X \mid W)\cdot P(W), \qquad (1)$$

где Γ — множество всех возможных последовательностей слов, $P(W|X)$ — условная вероятность появления последовательности акустических векторов X для заданной последовательности слов W, а $P(W)$ — априорная вероятность появления последовательности слов W. Выражение $P(W|X)$ обычно называют акустико-фонетической моделью, а $P(W)$ — моделью языка [1]. Типичная блок-схема системы АРР, построенной на этих принципах, представлена на **рис. 2**.

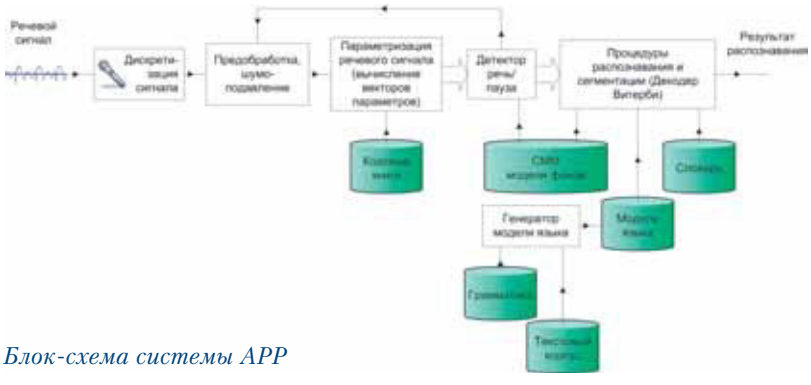

Рис. 2. Блок-схема системы АРР

Дискретизация речевого сигнала

На первом этапе необходимо произвести дискретизацию речевого сигнала, т. е. выполнить преобразование из аналогового в цифровой вид. Существует довольно много разнообразных форматов представления дискретного сигнала. Наиболее распространённой считается импульсно-кодовая модуляция (Pulse Code Modulation, PCM) с длиной слова 16 бит. В телефонии часто используются аудиокомпандеры с μ-законом (μ-Law) в Северной Америке и Японии или A-законом (A-Law), в Европе и остальном мире.

Предобработка и шумоподавление

Наличие высокого уровня фонового шума может ощутимо влиять на качество распознавания и привести к значительному увеличению числа ошибок распознавания. Поскольку система АРР может быть использована в местах с высоким уровнем фонового шума, например, в людном месте, в аэропорту, в движущейся машине, на улице, необходимо выполнять шумоподавление, задача которого состоит в полном или частичном удалении аддитивного шума.

Разработано довольно много различных методов удаления аддитивного некоррелированного шума, которые успешно используются в системах распознавания речи [2]. Наиболее распространены два метода: спектральное вычитание [3–5] и винеровская фильтрация [6–8]. Оба этих метода основаны на оценке характеристик фонового шума в моменты отсутствия речи.

Несмотря на простоту реализации и довольно высокое качество шумоподавления, спектральное вычитание менее популярно в системах АРР, так как обладает одним серьёзным недостатком — генерацией искусственных музыкальных тонов (музыкальным шумом) [9], которые негативно влияют на качество распознавания.

На сегодняшний день применение винеровского фильтра в качестве препроцессора для системы АРР стандартизировано ETSI [10].

Параметризация речевого сигнала

Основная задача параметризации заключается в вычислении адекватного параметрического представления речевого сигнала. В качестве параметров наибольшее распространение получило представление речевого сигнала, основанное на гребёнке полосовых кепстральных фильтров, распределённых по шкале мела. Такое представление хорошо зарекомендовало себя в системах распознавания речи, так как такие параметры значительно снижают размерность представления речевого сигнала, достаточно информативны и адекватно моделируют источник речевого сигнала. Алгоритм вычисления векторов параметров речевого сигнала, блок-схема которого представлена на **рис. 3**, состоит из последовательности следующих операций:

1. Коррекция верхних частот, которая состоит в обработке речевого сигнала фильтром с передаточной характеристикой [11]

$$H(z) = 1 - \alpha z^{-1} \qquad (2)$$

где $\alpha \leq 1$. Фильтр практически вычисляет первую производную сигнала. Физический смысл этой операции состоит в удалении постоянной составляющей и усилении высокочастотной части спектра, которая у речевого сигнала в среднем имеет затухание 6 дБ/декаду.

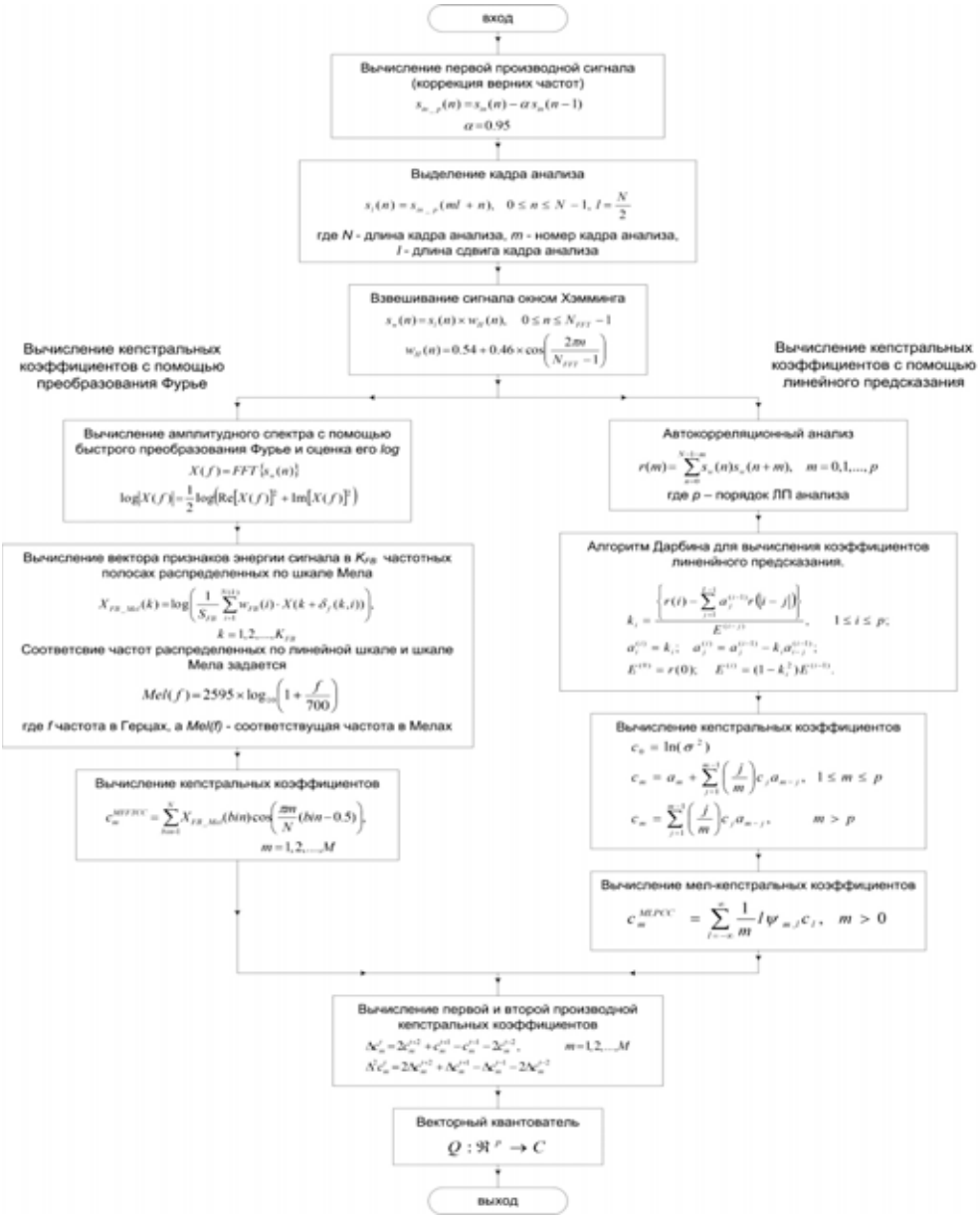

Рис. 3. Алгоритм вычисления параметров речевого сигнала

2. Сегментация на фреймы длиной 20–40 мс., на которых сигнал считается стационарным и его параметры не изменяются. Фреймы берутся с перекрытием

$$s_l(n) = s_{in\_p}(ml + n), \quad 0 \le n \le N-1, \; l = \frac{N}{2}, \tag{3}$$

где N — длина фрейма анализа, m — номер фрейма, а l — длина сдвига фрейма по сигналу.

3. Взвешивание окном. Каждый фрейм взвешивается окном для уменьшения краевых эффектов. Для речевого сигнала обычно используется окно Хемминга

$$s_w(n) = s_l(n) \times w_H(n)$$
$$w_H(n) = 0.54 + 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \qquad (4)$$

Далее для вычисления значений выходов гребёнки полосовых кепстральных фильтров используют либо преобразование Фурье, либо анализ на основе линейного предсказания.

Вычисление с помощью преобразования Фурье

a.4. Вычисление амплитудного спектра сигнала

$$X_{lin}(f) = FFT\{s_w(n)\}$$
$$\log\left(\left|X_{lin}(f)\right|\right) = \frac{1}{2}\log\left(\mathrm{Re}\left[X_{lin}(f)\right]^2 + \mathrm{Im}\left[X_{lin}(f)\right]^2\right). \qquad (5a)$$

a.5. Вычисление отклика гребёнки треугольных полосовых фильтров, распределённых по шкале мела [1, 12]. Это преобразование основано на имитации структуры критических частотных полос человеческого уха

$$X_{FB\_Mel}(k) = \log\left(\frac{1}{S_{FB}}\sum_{i=1}^{N(k)} w_{FB}(i) \cdot X_{FB\_lin}(k + \delta_f(k,i))\right), \qquad (6a)$$
$$k = 1, 2, ..., K_{FB}.$$

Шкала мела представляет собой нелинейное преобразование линейной частотной шкалы в соответствии с формулой

$$Mel(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \qquad (7a)$$

a.6. Вычисляем кепстральные коэффициенты

$$\log X(\omega) = \sum_{n=-\infty}^{\infty} c(n) e^{j\omega n}, \qquad (8a)$$

которые представляют собой обратное преобразование Фурье от логарифма, полученного на предыдущем отклика гребёнки фильтров [13]. Для вычисления кепстральных коэффициентов используется дискретное косинусное преобразование [12]

$$c_m^{MFFTCC} = \sum_{i=1}^{N} X_{FB\_Mel}(i) \cos\left(\frac{\pi m}{N}(i - 0.5)\right), \quad m = 1, 2, ..., M \qquad (9a)$$

В итоге получается множество мел-кепстральных коэффициентов.

Вычисление с помощью анализа на основе линейного предсказания

b.4. Автокорреляционный анализ

$$r(m) = \sum_{n=0}^{N-1-m} s_w(n) s_w(n+m), \quad m = 0, 1, ..., p \qquad (5b)$$

где S_w — взвешенный окном Хемминга (4) сигнал, p — порядок модели линейного предсказания.

b.5. Вычисляются коэффициенты линейного предсказания и коэффициенты отражения на основе автокорреляционных коэффициентов. Для вычислений используется метод, который известен как алгоритм Дарбина [14]

$$k_i = \frac{\left\{ r(i) - \sum_{j=1}^{L-1} a_j^{(i-1)} r\left(|i-j|\right) \right\}}{E^{(i-j)}}, \quad 1 \le i \le p; \tag{6b}$$
$$a_i^{(i)} = k_i; \quad a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)};$$
$$E^{(0)} = r(0); \quad E^{(i)} = (1-k_i^2)E^{(i-1)},$$

где $a_m = a_m^{(p)}, 1 \le m \le p$ — множество коэффициентов линейного предсказания, а k_i — множество коэффициентов отражения. Уравнения (6b) решаются рекурсивно для $i = 1,2,..,p$ — порядок модели линейного предсказания.

b.6. Вычисляем кепстральные коэффициенты на основе полученных коэффициентов линейного предсказания [1]

$$c_0 = \ln(\sigma^2)$$
$$c_m = a_m + \sum_{j=1}^{m-1} \left(\frac{j}{m}\right) c_j a_{m-j}, \quad 1 \le m \le p, \tag{7b}$$
$$c_m = \sum_{j=1}^{m-1} \left(\frac{j}{m}\right) c_j a_{m-j}, \qquad m > p$$

где σ^2 коэффициет усиления в модели линейного предсказания.

b.7. Вычисление кепстральных коэффициентов, распределённых по шкале мела, выполняется билинейным преобразованием частотной оси, полученных кепстральных коэффициентов [15], которое состоит в фильтрации последовательностью фазовых фильтров

$$c_m^{MLPCC} = \sum_{l=-\infty}^{\infty} \frac{1}{m} l \psi_{m,l} c_l, \quad m > 0 \tag{8b}$$

где импульсный отклик фильтра с передаточной характеристикой

$$H_n(z) = \frac{(1-\alpha^2)z^{-1}}{(1-\alpha z^{-1})^2} \left[\frac{z^{-1}-\alpha}{1-\alpha z^{-1}} \right]^{n-1}, \quad n > 0 \tag{9b}$$

Степень частотной деформации управляется изменением коэффициента α. Например, значение α для речевого сигнала с частотой дискретизации 8 кГц равно 0,3624 [16].

В итоге получается множество мел-кепстральных коэффициентов.

5. Последним параметром является логарифм энергии сигнала, вычисленный на фрейме анализа

$$E = 10 \log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1} s_w^2(n)\right) \qquad (10)$$

6. Для учёта динамических характеристик сигнала используют коэффициенты ортогонального полинома первого и второго порядка, полученные из регрессионного анализа каждого кепстрального коэффициента, и энергии сигнала, которые рассматриваются как функция времени c_m^t и E^t [17, 18]

$$\Delta c_m^t = \frac{\sum_{k=-K}^{K} w_k c_m^{t+k}}{\sum_{k=-K}^{K} w_k^2}. \qquad (10)$$

Эти коэффициенты являются оценкой первой и второй производных кепстральных коэффициентов и энергии по времени [19].

7. Полученные векторы параметров поступают на вход кодера (векторного квантователя), который выполняет отображение p-мерного непрерывного пространства $\Re^p$ в конечное множество векторов C

$$Q : \Re^p \to C, \qquad (11)$$

т.е. преобразование непрерывных значений векторов параметров в последовательность символов из конечного алфавита с помощью кодовых книг.

Детектор речь/пауза

Детектор речь/пауза — очень важная часть системы АРР. Основная задача детектора состоит в отбрасывании фреймов, не содержащих полезный речевой сигнал, что позволяет сократить число ошибок, связанных со вставками, и снизить вычислительную нагрузку на блок распознавания, так как в этом случае он работает только на полезном сигнале. Кроме того, детектор управляет шумоподавителем, указывая участки с отсутствием речевого сигнала, т.е. участки, где можно производить оценку статистики фонового шума, которая необходима для работы одноканального алгоритма шумоподавления, например, такого, как спектральное вычитание или винеровская фильтрация.

Существует большое количество различных алгоритмов детекторов, например можно использовать детектор, основанный на технологии скрытой марковской модели (СММ) [20]. Этот алгоритм использует одну модель для шума и одну модель для речи, при этом происходит анализ каждого фрейма входного сигнала и вычисление правдоподобия для каждой из этих моделей. В момент, когда правдоподобие речевой модели превышает некоторый порог правдоподобия, принимается решение о наличии речевого сигнала. Такой алгоритм демонстрирует приемлемое качество при вычислительной сложности, которая позволяет использовать его в системах реального времени.

Распознавание и сегментация

Блок распознавания и сегментации (декодер Витерби) выполняет распознавание речевого высказывания. Этот блок построен на основе скрытых марковских моделей (СММ),

которые представляют собой статистический метод сравнения распознаваемого высказывания с эталонами, используя при этом акустические модели слов и модель языка. Это одна из наиболее популярных на сегодняшний день моделей, используемых в системах распознавания речи, так как СММ обеспечивает хорошее представление речевого сигнала, что позволяет достичь довольно высокой точности распознавания.

СММ состоит из марковской цепи с конечным числом состояний, которая моделирует временные изменения сигнала, и конечного множества выходных распределений вероятностей, которые позволяют моделировать спектральные вариации сигнала. В качестве базовых единиц системы часто используются фоны, для каждого из которых создана отдельная СММ, при этом произнесение каждого фона описывается последовательностью векторов спектральных характеристик сигнала. С лингвистической точки зрения фоны соответствуют фонемам. Марковская модель слова из словаря системы получается конкатенацией элементарных моделей фонов.

В процессе распознавания неизвестного высказывания необходимо найти наиболее подходящую модель M_i, которая максимизировала бы вероятность $P(M_i|X,\Theta)$ при фиксированном множестве параметров модели Θ и наблюдаемой в данный момент последовательности векторов спектральных характеристик речевого сигнала X. При этом результатом распознавания высказывания X будет слово, связанное с моделью M_i такое, что

$$i = \arg\max_j P(M_j \mid X, \Theta). \qquad (12)$$

Метод нахождения наилучшей модели основан на динамическом программировании и называется алгоритмом Витерби [21].

Генератор модели языка

Блок генератора модели языка необходим для генерации модели языка на основе грамматики и текстового корпуса. Использование модели языка позволяет сократить число гипотез, появляющихся в процессе перебора и рассматривать только допустимые для данного языка.

Обычно архитектура системы АРР строится в соответствии с приведённой блок-схемой (рис. 2) и имеет довольно строгое разделение на перечисленные блоки, что обеспечивает достаточную гибкость для реализации системы в удалённом варианте.

Реализация удалённой системы АРР

По месту обработки речевого сигнала: на клиентском компьютере, на сервере или на обоих, определяют архитектуру системы АРР. Первая — наиболее привычная и распространённая — называется *встроенной*. При такой реализации вся обработка речевого сигнала и распознавание выполняются на клиентском компьютере или мобильном устройстве, т.е. такая система АРР полностью находится на компьютере или устройстве пользователя [22–23].

Такое решение обладает целым рядом недостатков. Прежде всего, это проблемы, возникающие при разработке систем APP и связанные с большим разнообразием в существующих архитектурах компьютеров и мобильных устройств, поскольку необходимо выполнить портирование системы под каждую вычислительную архитектуру. Для мобильных устройств это ещё и ограничения, накладываемые на вычислительные ресурсы:

- невысокая вычислительная мощность;
- часто только целочисленная арифметика;
- небольшой объём и низкая скорость оперативной памяти;
- небольшой объём и низкая скорость доступа энергонезависимой памяти.

Поэтому для разработки встроенных систем APP для мобильных устройств требуется глубокая оптимизация программ с целью снижения требований к вычислительным ресурсам, что приводит к значительному росту трудоёмкости, а следовательно, удорожанию и увеличению времени разработки. Кроме того, возникают большие трудности с поддержкой и обновлением программного обеспечения пользовательского компьютера, а для мобильных устройств обновление часто вообще невозможно. Правда, стремительный рост производительности процессоров и снижение стоимости памяти дают определённую надежду на преодоление этих проблем. А большое число работ, которые ведутся в этом направлении [23–24], без сомнения, в скором времени позволят разработать оптимальные и устойчивые встроенные системы APP, которые необходимы в ограниченных задачах, например, голосовой набор телефонного номера, голосовое управление объектом или голосовое управление в автомобиле, где встроенная система — единственно возможная архитектура.

Второй и третий варианты построения систем APP, которые называют удалёнными, — более мощная и более гибкая альтернатива встроенной системе. В такой архитектуре клиентский компьютер (мобильный телефон, смартфон, КПК или нетбук) осуществляет ввод и передачу речевого сигнала или вычисленных параметров по цифровому каналу связи на удалённый сервер, который выполняет распознавание полученной последовательности данных. У такой архитектуры отсутствуют ограничения на вычислительные ресурсы клиентского компьютера, что даёт возможность использовать более современные и более сложные алгоритмы распознавания, кроме того, возможна централизованная поддержка и обновление серверной программы системы APP. К сожалению, такая архитектура является источником дополнительных проблем для системы распознавания. Во-первых — это акустический шум, который может добавляться к речевому сигналу, поскольку теперь пользователь имеет возможность находиться в очень шумном окружении, например, гостинице, аэропорту, вокзале и т.д. Как следствие, высокий уровень фонового шума может привести к серьёзному снижению качества распознавания. Во-вторых, появляются помехи, вносимые цифровым каналом передачи данных, которые проявляются в разрушении передаваемых данных или потере целого пакета. В основном это связано либо с ухудшением качества или вообще потерей связи (характерно для беспроводных сетей), либо с уменьшением пропускной способности при передаче по IP сетям. Кроме того, такая архитектура характеризуется дополнительной нагрузкой на сетевые ресурсы и повышенными требованиями к пропускной способности канала связи.

Возможны два способа реализации удалённой системы распознавания речи (УСРР). Первый — *сетевая система распознавания речи* (ССРР), структурная схема которой приведена на **рис. 4**. В этом случае клиентский компьютер обеспечивает ввод речевого сигнала, который далее кодируется стандартным речевым кодеком и передаётся на сервер, где происходит вычисление параметров речевого сигнала и распознавание.

Рис. 4. Сетевая система распознавания речи

Второй способ реализации удалённой системы распознавания — распределённая система распознавания речи (РСРР), структурная схема которой приведена на **рис. 5**. При такой реализации системы АРР параметризация речевого сигнала переносится на клиентскую сторону и в канал связи передаётся уже вычисленная последовательность векторов параметров, а на серверной стороне выполняется только распознавание этой последовательности. При такой реализации снижается требование к пропускной способности канала связи, так как параметризация речевого сигнала значительно снижает его размерность и избыточность.

Рис. 5. Распределённая система распознавания речи

Сетевая система распознавания речи

Структурная схема ССРР приведена на **рис. 6**. Как видно из схемы, на клиентском компьютере последовательно выполняются: кодирование сигнала речевым кодеком, канальное кодирование полученного битового потока и упаковка в пакеты, в виде которых речевой сигнал передаётся по цифровому каналу связи, например, по Интернету. На сервере принятые пакеты распаковываются и канально декодируются, т.е. происходит дешифрация и восстановление

Рис. 6. Структурная схема ССРР

полученного битового потока, если произошли ошибки при передаче. Далее происходит параметризация восстановленного речевого сигнала и после этого вычисленная последовательность векторов параметров попадает в блок распознавания (декодер Витерби).

Кодирование речевого сигнала

Одно из основных отличий CCPP — использование стандартных речевых кодеков, основная цель которого — снижение размерности передаваемого по цифровым каналам речевого сигнала. Стремясь повысить качество сигнала и снизить скорость передачи, разработчики постоянно совершенствуют существующие и разрабатывают новые стандарты кодирования речевых сигналов, поэтому число речевых кодеков неуклонно растёт. Более подробно с обзором стандартов речевого кодирования можно познакомиться в работе [25] или на веб-сайте международных организаций ITU (www.itu.int), ETSI (www.etsi.org).

Наиболее популярны в системах передачи речи по IP сетям (протоколы Voice over IP, VoIP), кодеки G.723.1, G.729 и G.711. Они также используются как часть стандартов ITU H.323 и IETF SIP. Ещё один широко распространённый кодек GSM в основном используется для передачи речевых данных в беспроводных сетях, например, в мобильной телефонии, но может использоваться и для передачи в IP сетях. Далее приведены краткие описания нескольких самых распространённых речевых кодеков.

GSM

Кодек GSM (Global System for Mobile Communication) — первый цифровой стандарт кодирования речи, использованный в телефонах GSM. Внутри этого стандарта существует несколько разновидностей кодеков Full Rate (FR), Half Rate (HR), Enhanced Full Rate (EFR) и Adaptive Multi Rate (AMR). Все они соответствуют европейским телекоммуникационным стандартам. Входным сигналом для этих кодеков является 13-битный PCM речевой сигнал с частотой дискретизации 8 кГц. Кодирование выполняется на фрейме сигнала длиной 20 мс. (160 отсчётов). Фреймы берутся последовательно без перекрытия.

Full Rate (FR) [26] был первым стандартом (стандартизирован в 1987 году) цифрового кодирования речи, используемым в цифровой системе мобильной связи GSM. Кодек основан на принципе долговременного линейного предсказания с регулярным импульсным возбуждением (Regular Pulse Excitation — Long Term Prediction (RPE-LTP)). Скорость передачи данных этого кодека составляет 13 кбит/с. Качество закодированной речи довольно плохое по современным меркам, но во время разработки (начало 90-х годов прошлого века) это был хороший компромисс между вычислительной сложностью и качеством. Входной фрейм речевого сигнала длиной 160 отсчётов кодируется в блок длиной 260 бит, что соответствует скорости передачи 13 кбит/с. При декодировании блок 260 бит восстанавливается в 160 отсчётов. Скорость по стандарту 22.8 кбит/с. Разница в 9.8 кбит/с. используется для помехозащищающих кодов. Кодек описывается спецификацией GSM 06.10.

Half Rate (HR) — система кодирования речи для GSM, разработанная в начале 90-х годов прошлого века компанией Motorola для увеличения пропускной способности канала передачи. Скорость передачи составляет 5.6 кбит/с, т. е. он требует половины полосы пропускания кодека Full Rate. Кодек основан на принципе (Code Excited Linear Prediction — Vector Sum Excited Linear Prediction) (CELP-VSELP). Реальная скорость передачи 11.4 кбит/с. Качество распознавания для обоих кодеков сравнимо. Кодек описывается спецификацией GSM 06.20.

Enhanced Full Rate (EFR) [27] — стандарт кодирования речи был разработан с целью улучшения качества кодека FR. Скорость передачи EFR кодека 12.2 кбит/с для кодирования речи и 10.6 кбит/с для корректирования ошибок. EFR обеспечивает качество, сравнимое с проводной связью, как в бесшумных условиях, так и при фоновых шумах. Стандарт кодирования речи EFR сравним с наилучшим режимом AMR. Схема кодирования речи основана на ACELP (Algebraic Code Excited Linear Prediction) алгоритме. Кодек описывается спецификацией GSM 06.60.

Adaptive Multi-Rate (AMR) [28] — последний стандарт кодирования речевого сигнала для GSM. В октябре 1998 г. ассоциация 3GPP приняла AMR как стандартный речевой кодек, и сейчас он широко используется в GSM и IP сетях. AMR обладает широким набором скоростей кодирования/декодирования речи и позволяет гибко переключаться в различные режимы в зависимости от текущих условий приёма и ёмкости сети. При ухудшении качества радиосигнала или пропускной способности автоматически выбирается режим с более низкой скоростью передачи данных. Это улучшает качество и надёжность соединения, незначительно жертвуя качеством речевого сигнала (при соотношении сигнал/шум = 4–6 дБ). Кодек может работать в 14 режимах, что позволяет динамически изменять скорость потока данных от 4.5 до 12.2 кбит/с. Кодек описывается спецификацией GSM 06.74.

ITU кодеки

G.723.1 — один из базовых кодеков для приложений IP-телефонии, стандартизирован ITU-T в рекомендации G.723.1 в ноябре 1995 года. Кодек имеет два режима работы: 6.4 кбит/с (фрейм имеет размер 189 битов, дополненных до 24 байтов) и 5.3 кбит/с (фрейм имеет размер 158 битов, дополненных до 20 байтов). Режим работы может меняться динамически от фрейма к фрейму. Оба режима обязательны для реализации. Кодек G.723.1 имеет встроенный детектор речевой активности (VAD) и обеспечивает генерацию комфортного шума на удалённом конце в период молчания. Эти функции специфицированы в приложении A (Annex A) к рекомендации G.723.1. Параметры фонового шума кодируются очень маленькими кадрами размером 4 байта. Если параметры шума не меняются существенно, передача полностью прекращается. Кодек построен на основе линейного предсказания.

G.729 — также один из основных кодеков IP-телефонии, стандартизирован ITU-T в рекомендации G.729. Скорость передачи кодека 8 кбит/с. Алгоритм кодека основан на модели кодирования, которая использует алгоритм линейного предсказания с возбуждением по алгебраической кодовой книге (Conjugate Structure — Algebraic Code Excited Linear Prediction, CS-ACELP). Кодер оперирует фреймами речевого сигнала длиной 10 мс, с частотой квантования 8 КГц, что соответствует 80 16-битным PCM отсчётам. Для каждого фрейма речевого сигнала вычисляются параметры модели (коэффициенты фильтра линейного предсказания, индексы и коэффициенты усиления в адаптивной и фиксированной кодовых книгах). Далее эти параметры кодируются и передаются в канал. Кодек G.729 A аналогичен G.729, но стандартизован для целочисленных операций.

G.711 — это самый простой кодек, который является ITU-T стандартом для аудиокомпандирования. Впервые был представлен в 1972 г. Кодек компандирует

каждый 16-битный отсчёт входного сигнала в 8-битный с частотой квантования сигнала 8000 Гц. Таким образом, G.711 кодек создаёт поток 64 кбит/с. Существуют два основных алгоритма, представленных в стандарте, μ-law (используется в Северной Америке и Японии) и A-law (используется в Европе и в остальном мире). Оба алгоритма логарифмические, но более поздний A-law был изначально предназначен для компьютерной обработки процессов.

Реальная скорость передачи в VoIP каналах несколько больше, чем позволяет кодек. Например, кодек G.729 обеспечивает скорость передачи 8 кбит/с, но реальная скорость передачи по IP протоколу 32.2 кбит/с. Это увеличение происходит из-за добавления заголовков при использовании различных протоколов, т.е. заголовок Real Time Protocol (RTP) составляет 12 байт, заголовок User Datagram Protocol (UDP) — 8 байт, заголовок IP протокола — 20 байт и, наконец, заголовок Ethernet протокола добавляет 18 байт. Итого 58 дополнительных байт на каждый пакет.

Вычисление параметров речевого сигнала

Процесс кодирования/декодирования речевым кодеком вносит искажения в речевой сигнал, которые оказывают негативное влияние на качество распознавания. Одна из первых работ [29], посвящённая исследованию влияния кодирования речи на точность распознавания, показала, что существует две основные причины, снижающие точность распознавания.

Во-первых, речевые базы данных сигналов, которые используются для обучения системы АРР, содержат речевые сигналы, отличающиеся от сигналов, обработанных речевыми кодеками, т.е. были закодированы и декодированы. Один из способов снизить влияние этого фактора — выполнить обучение СММ на речевом сигнале, предварительно обработанном речевым кодеком. Такие эксперименты были успешно проведены при разработке системы распознания речевых сигналов, записанных с сотовых телефонов, и описаны в [30]. Однако большое число различных стандартов кодирования, используемых в современных системах передачи речи, значительно усложняет решение этой задачи таким способом, так как необходимо выполнять обучение для каждого кодека. Кроме того, в больших распределённых сетях часто последовательно используется несколько различных типов кодеков.

Во-вторых, при сжатии ухудшается качество речевого сигнала с точки зрения системы распознавания [29, 31]. Связано это с тем, что параметризация речевого сигнала для кодирования и для распознавания отличается, поскольку основным критерием при разработке речевого кодека является повышение субъективного качества восприятия речи, а также сохранение эмоциональной и тембральной окраски речевого высказывания. Такой критерий, в общем, противоречит объективной оценке речевых параметров для системы АРР. Наиболее распространены кодеки, основанные на линейной авторегрессионной модели процесса формирования речи, которая моделирует спектральную огибающую отклика речевого тракта, используя коэффициенты линейного предсказания, а в системах распознавания широко распространены мел-кепстральные коэффициенты, моделирующие процесс восприятия речи человеком.

Одним из решений этой проблемы может быть вычисление параметров речевого сигнала непосредственно из битового потока речевого кодека, не используя декодер, т.е. репараметризация или транскодирование параметров без восстановления речевого сигнала. Работа [32] продемонстрировала повышение качества работы системы

распознавания при использовании такого способа вычисления параметров. Причиной такого улучшения является, прежде всего, то, что транскодирование исключает искажения, которые вносит декодер при восстановлении сигнала. Ниже приведён пример такого транскодирования, а именно вычисление мелкепстральных параметров прямо из линейных спектральных пар, которые широко применяются для кодирования сигнала речевыми кодеками [33].

Для заданного множества линейных спектральных пар порядка M, $\Omega = \{\omega_1, \omega_2, ..., \omega_M\}$ коэффициенты линейного предсказания могут быть получены, используя следующее выражение:

$$A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i} = \frac{(P(z)-1) + (Q(z)-1)}{2}, \tag{13}$$

где

$$\begin{aligned} P(z) - 1 &= \prod_{i=2,4,...}^{M} (1 + x_i z^{-1} + z^{-2}) - z^{-1} \prod_{i=2,4,...}^{M} (1 + x_i z^{-1} + z^{-2}) - 1 \\ &= z^{-1} \sum_{i=2,4,...}^{M} \left[(x_i + z^{-1}) \prod_{j=0,2,4...}^{i-2} (1 + x_j z^{-1} + z^{-2}) \right] - z^{-1} \prod_{i=2,4,...}^{M} (1 + x_i z^{-1} + z^{-2}) \end{aligned} \tag{14}$$

а

$$\begin{aligned} Q(z) - 1 &= (1 + z^{-1}) \prod_{i=1,3,...}^{M-1} (1 + x_i z^{-1} + z^{-2}) - 1 \\ &= \prod_{i=1,3,...}^{M-1} (1 + x_i z^{-1} + z^{-2}) + z^{-1} \prod_{i=1,3,...}^{M-1} (1 + x_i z^{-1} + z^{-2}) - 1 \\ &= z^{-1} \sum_{i=1,3,...}^{M} \left[(x_i + z^{-1}) \prod_{j=-1,1,3...}^{i-2} (1 + x_j z^{-1} + z^{-2}) \right] - z^{-1} \prod_{i=1,3,...}^{M-1} (1 + x_i z^{-1} + z^{-2}) \end{aligned} \tag{15}$$

при $x_i = -2\cos\omega_i$ для $i = 1,2,...,M$, и $x_{-1} = -z^{-1}$ так, что $1 + x_{-1} z^{-1} + z^{-2} = 1$. Далее мел-кепстральные параметры получаются, используя формулы (7b) и (8b).

Ещё одним небольшим плюсом такого подхода является снижение вычислительной нагрузки, поскольку нет необходимости восстанавливать речевой сигнал.

Однако этот подход обладает тем же недостатком — существующее разнообразие кодеков и, следовательно, необходимость разработки алгоритмов транскодирования для каждого стандартного кодека, который используется для передачи речевого сигнала. Кроме того, часто длина фрейма и частота, с которой вычисляются параметры, могут не совпадать для кодека и системы распознавания. Правда, это несоответствие может быть ослаблено либо интерполяцией [34], либо уменьшением числа СММ состояний [35].

Тем не менее, на сегодняшний день многие из алгоритмов транскодирования уже разработаны, опубликованы и успешно применяются, например, для кодека GSM RPE-LTP [36–38] или для кодека ITU-T G.723.1 [32].

Распределённая система APP

Второй вариант реализации УСРР называется распределённой системой распознавания речи (РСРР), структурная схема которой приведена на **рис. 7**. Как видно из схемы, принципиальное отличие такой реализации от ССРР состоит в том, что параметризация речевого сигнала перенесена с сервера на клиентский компьютер или устройство. Таким образов, сервер получает уже вычисленную последовательность векторов параметров, которую необходимо распознать. Такая реализация обладает явным преимуществом перед ССРР, так как исключает процесс кодирования/декодирования стандартным речевым кодеком, который является одним из источником снижения точности системы APP. Кроме того, передача речевых параметром снижает требования к пропускной способности канала передачи данных. Однако перенос вычислений на клиентский компьютер требует установки дополнительного программного обеспечения, поскольку параметризация остаётся специфической процедурой практически для каждой системы APP. Для некоторых мобильных устройств такая установка практически невозможна. Утешает, что таких устройств становится всё меньше и меньше.

Рис. 7. Структурная схема РСРР

Стандарты РСРР

На сегодняшний день рабочей группой STQ Aurora DSR в ETSI разработан и опубликован целый ряд стандартов регламентирующих построение распределённых систем автоматического распознавания речи [39]. Список этих стандартов приведён в таблице 1.

Стандарты, разработанные в ETSI STQ-Aurora, регламентируют алгоритмы предобработки, шумоподавления, параметризации речевого сигнала и канального кодирования. В качестве основного набора параметров спецификации рекомендуют мел-кепстральные параметры, которые широко и успешно используются в современных системах распознавания речи. Первый стандарт ES 201108, который был принят в 2000 году, определил набор параметров для представления речевого сигнала и требуемую скорость передачи в 4,8 кбит/с.

Для мобильных устройств, которые часто могут использоваться в очень шумном окружении, был разработан стандарт вычисления параметров более устойчивый к шуму, основная цель разработки которого — повышение устойчивости к шумам и, соответственно, снижение числа ошибок при использовании УСРР в среде с высоким уровнем фонового шума. В качестве кандидатов было проверено несколько алгоритмов шумоподавления.

Таблица 1

Список стандартов для PCPP

№ стандарта	Описание	Разработчик
ES 201 108 [40]	Параметризация речевого сигнала, мел-кепстральные параметры	ETSI STQ-Aurora
ES 202 050 [41]	Улучшенное вычисление параметров речевого сигнала	ETSI STQ-Aurora
ES 202 211 [42]	Расширенная параметризация речевого сигнала	ETSI STQ-Aurora
ES 202 212 [43]	Расширенное, улучшенное вычисление параметров	ETSI STQ-Aurora
TS 26.243 [44]	Спецификация на использование вычислений с фиксированной точкой для стандартов ES 202 050 и ES 202 212	3GPP
RFC3557	Использование протокола RTP для стандарта ES 201 108	IETF
RFC4060	Использование протокола RTP для стандартов ES 202 050, ES 202 211 и ES 202 212	IETF

Испытания проводились на различных речевых базах данных. Результатом этой работы стал стандарт ES 202050, в котором основным алгоритмом шумоподавления выбран двухпроходный винеровский фильтр.

Позднее рабочая группа ETSI STQ-Aurora расширила стандарты с целью включения в них возможности восстановления речевого сигнала из последовательности векторов параметров, а также включила поддержку тональных языков. Результатом стала публикация стандартов ES 202211 и ES 202212, которые являются расширением стандартов ES 201108 и ES 202050 соответственно.

Все перечисленные стандарты ETSI STQ-Aurora содержат спецификации алгоритмов на языке «C» с использованием плавающей точки.

Одновременно в IETF разрабатывались спецификации, определяющие передачу мультимедийных данных по компьютерным IP сетям. Эти протоколы были объединены под общим названием Voice over IP (VoIP). Основным транспортным протоколом для передачи аудио-, видеоданных был определён протокол Real-time Transport Protocol (RTP). Чтобы объединить работы и обеспечить возможность использовать RTP протокол для передачи последовательности параметров речевого сигнала, в PCPP были разработаны две дополнительные спецификации RFC3557 для стандарта ES 201108 и RFC4060 для стандартов ES 202050, ES 202211 и ES 202212.

3 GPP совместно с IBM и Nuance провели обширное тестировании различных систем распознавания на различных речевых базах данных. В результате этого тестирования для применения в системах распознавания было рекомендовано представление речевого сигнала, специфицированное стандартами ES 202050 и ES 202212. Кроме того, 3 GPP разработала спецификацию TS 26.243, позволяющую использовать целочисленную арифметику для этих стандартов.

Повышение устойчивости УСРР к ошибкам канала связи

Как уже отмечалось выше, цифровой канал передачи данных при реализации УСРР служит дополнительным источником ошибок, приводящих к снижению качества распознавания. Эти ошибки возникают из-за ухудшения качества связи, которое для беспроводных сетей связано с наличием помех в канале связи или ухудшением условий приёма/передачи, а для IP сетей — с недостаточной пропускной способностью канала. Возникающие ошибки бывают двух типов: либо искажение информации внутри пакета, либо потеря целого пакета. Для повышения устойчивости УСРР, т.е. снижения зависимости качества распознавания от наличия таких ошибок, используют канальное кодирование [45] и маскирование ошибок [46]. Эти методы могут использоваться совместно.

Канальное кодирование

Основная цель канального кодирования (КК) — защита информации от возникающих ошибок в канале. Методы КК, которые применяются в УСРР, можно разделить, как показано на **рис. 8** на две основные группы. Первый — это группа методов, которые представляют собой помехозащитное кодирование, а именно исправление ошибок методом упреждения (Forward Error Correction, FEC). Такое кодирование основано на введении искусственной избыточности в передаваемые данные. Эти методы включают методы, независимые от передаваемых данных, методы, зависимые от передаваемых данных и коды с неравномерной защитой символов. Вторая группа — это перемежение (interleaving), который заключается в переупорядочивании информации во время передачи для снижения влияния пачек ошибок, т.е. нескольких возникающих подряд ошибок. Описываемые методы в большей мере относятся к РСРР, так как в ССРР в основном используются методы защиты, регламентированные стандартами используемых речевых кодеков (GSM, G.723 и т.д.). Однако их вполне можно использовать и в ССРР [47].

Рис. 8. Методы канального кодирования

Методы, независимые от данных, представляют собой классическое помехоустойчивое кодирование передаваемой информации. При этом в УСРР успешно используются самые различные типы кодирования:

- линейные блочные коды [48];
- циклические коды [49] [43] или их разновидность коды Рида-Соломона [50];
- свёрточные коды, а именно алгоритм Витерби с мягким выходом (soft-output Viterbi algorithm, SOVA) [51] или Max-Log-MAP алгоритм [52].

Методы, зависимые от данных, основаны на дублировании передаваемой информации в разных пакетах перед передачей в канал связи. Тогда при возникновении ошибки или потере пакета утраченные данные могут быть восстановлены. В [53] предложен алгоритм, совместимый с ESTI стандартом [40].

Коды с неравномерной защитой символов также можно отнести к методам, зависимым от передаваемых данных. Такое кодирование позволяет снизить избыточность передаваемых данных, так как позволяет изменять степень защиты в зависимости от важности данных для системы распознавания. Использование таких методов привлекательно в системах с низкой пропускной способностью [50, 54].

Перемежение (interleaving) фреймов, которое также широко используется в УСРР, применяется для снижения влияния пачек ошибок, поскольку такие ошибки сильнее сказываются на качестве распознавания, чем случайные ошибки. Смысл этого метода состоит в том, чтобы сделать последовательность ошибок более короткой. Перемежение может быть применено как на уровне векторов параметров, так и на уровне пакетов [55]. Пример такого перемежения показан на **рис. 9**. В этом примере по каналу связи было передано 16 фреймов параметров, упакованных по два фрейма в каждый пакет. В результате возникшей ошибки половина пакетов была потеряна. Если использовать нормальный порядок передачи фреймов, то потерянными оказались бы 8 последовательных фреймов. Использование перемежения позволяет сократить длину последовательных пропаданий до двух фреймов. Преимущество этого метода — отсутствие увеличения требуемого трафика. Однако возникает задержка, так как необходимо накопить несколько фреймов. Так, в примере на рис. 9 задержка составляет 12 фреймов. Перемежение более привлекательно в IP сетях, где каждый пакет содержит несколько последовательных фреймов речевых параметров. Более полное исследование различных способов перемежения для РСРР можно найти в [56].

Рис. 9. Перемежение пакетов

Перемежение можно использовать совместно с другими методами, зависимыми от передаваемых данных. В примере на **рис. 10** пакет содержит в себе помимо основного фрейма, который соответствует текущему моменту времени, ещё и дополнительный фрейм, соответствующий предыдущему или последующему моменту времени. Такие избыточные фреймы позволяют восстановить потерянные данные. В приведённом примере каждый пакет содержит четыре фрейма *и,* которые являются основными, и допол-

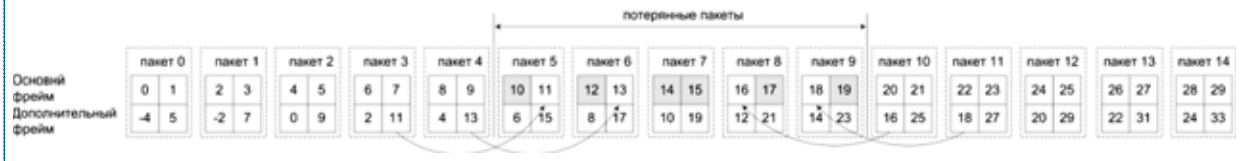

Рис. 10. Использование перемежения совместно с дублированием

нительные фреймы *и*. При этом задержка составляет 4 фрейма. Если при возникновении ошибки были потеряны пакеты с 5 по 9, то потерянными оказались бы фреймы с 10 по 19. Однако используемая схема коррекции ошибок позволяет не только сократить длину последовательности потерянных пакетов максимум до двух, но и восстановить фреймы 11, 13, 16 и 18.

Методы КК характеризуются различными возможностями по обнаружению и исправлению ошибок, возникающему запаздыванию реакции системы, вычислительной сложностью, объёмом избыточной информации, которую необходимо передавать. Все эти параметры необходимо учитывать при разработке системы.

Маскирование ошибок

Термин «маскирование ошибок» обозначает процедуру, цель которой — снижение или вообще устранение влияния на качество работы УСРР ошибок при передаче данных по каналу связи. Здесь имеются в виду ошибки, которые были обнаружены, но их не удалось исправить с помощью канального кодирования. Выполнять эту процедуру можно либо сразу после завершения канального декодирования, либо во время распознавания. В первом случае маскирование является независимой процедурой, а во втором необходимо вносить изменения в алгоритмы распознавания, и маскирование ошибок происходит совместно с распознаванием. На **рис. 11** приведена классификация методов маскирования ошибок.

Для независимого маскирования ошибок предложено несколько методов: вставка, интерполяция и статистическая оценка.

Рис. 11. Методы маскирования ошибок

- Вставка — самый простой метод маскирования, который заключается либо в замене пропущенного или испорченного фрейма речевых параметров фреймом с шумом, либо повторением последнего без ошибок фрейма [49].

- Интерполяция — это более сложный метод, который использует нахождение пропущенных параметров, используя интерполяционные полиномы. В основном используют линейную интерполяцию [57] или интерполяционные полиномы Эрмита [58].

- Метод статистической оценки аналогичен интерполяции, но вместо параметрической функции для генерации замещающих фреймов используется некоторая статистическая модель речевого сигнала [59].

Основная идея маскирования ошибок совместно с распознаванием аналогична методу статистической оценки и состоит в том, что можно использовать СММ системы распознавания, которая является мощной статистической моделью речевого сигнала. Для этого были разработаны различные алгоритмы, например взвешенный алгоритм Витерби и его модификации [59].

Сравнение сетевой и распределённой систем распознавания речи

Несмотря на то, что PCPP обладает рядом неоспоримых преимуществ перед CCPP:

- отсутствие стандартного речевого кодека, который искажает речевой сигнал и может вызвать снижение качества распознавания;
- ниже требования к пропускной способности каналов связи;
- возможность выполнять нестандартную предобработку и шумоподавление на клиентской части;
- унификация взаимодействия, т.е. в случае, когда речь комбинируется с другими видами ввода информации (клавиатура или touch screen), PCPP выглядит более органично, так как используется только канал передачи данных;
- выбор архитектуры не столь очевиден. Это связано прежде всего с тем, что реализация в реально существующих сетях и взаимодействие с существующими приложениями часто сводят на нет перечисленные преимущества PCPP.

Так, например, хотя использование стандартного речевого кодека вносит искажения в речевой сигнал, оно даёт возможность обратиться к сервису распознавания речи практически с любого существующего устройства (телефона, смартфона, ноутбука и т.д.) без его модификации, поскольку такие речевые кодеки присутствует практически во всех мобильных устройствах и операционных системах. Поэтому для организации взаимодействия с сервером APP пользователю не требуется установки никаких дополнительных программных продуктов на своё мобильное устройство или свой компьютер, т.е. для начала использования APP достаточно просто соединиться с сервером. Кроме того, при обновлении версии системы APP, соответственно, не требуется вносить изменения на клиентской стороне, что в случае с мобильным устройством часто бывает весьма проблематично. Правда, современные мобильные устройства значительно упрощают установку дополнительного программного обеспечения, что в будущем, видимо, облегчит использование PCPP.

Преимущество PCPP, связанное с пропускной способностью канала передачи данных, состоит в том, что CCPP требуется скорость передачи порядка 5–10 кбит/с для достижения удовлетворительного качества связи, тогда как PCPP требует всего 2 кбит/с или даже меньше. Однако в некоторых сетях конечная скорость передачи может оказаться одинаковой, например, передача данных при реализации CCPP и PCCP в GSM сетях будет осуществляться по каналам с одинаковой скоростью и в результате скорость передачи будет 22.8 кбит/с. в обоих случаях.

Считается, что устойчивость к акустическим помехам и ошибкам канала передачи у PCPP выше, чем CCPP. Однако следует принять во внимание, что в последнее время разработаны мощные методы шумоподавления и снижения влияния ошибок, возникающих в каналах передачи данных, которые позволяют компенсировать потери качества распознавания в обоих случаях как CCPP, так и PCPP.

И наконец, CCPP также обладает преимуществами, например, возможностью восстановления речевого сигнала. Если в системе требуется использовать восстановленный речевой сигнал, то CCPP оказывается более удобной.

К тому же, существует приложения, например, интерактивный речевой ответ (Interactive Voice Response, IVR), в которых CCPP — единственно возможный вариант организации взаимодействия. Однако надо принять во внимание тот факт, что речевой сигнал может быть восстановлен и из векторов параметров, если реализовать стандарты ESTI STQ-Aurora XFE и XAFE [42–43].

Более подробное сравнение можно найти в работах [39, 60–63].

Сетевые протоколы

Для построения UCPP используется набор протоколов, разработанных такими организациями, как The Internet Engineering Task Force (IETF, www.ietf.org), International Telecommunication Union (ITU, www.itu.int) и 3 rd Generation Partnership Project (3 GPP), которые объединяются под общим названием Voice over IP (VoIP). Эти протоколы позволяют решить задачу организации передачи голосовой и управляющей информации по цифровым коммуникационным сетям. Однако на сегодняшний день нет единого стандарта, регламентирующего передачу аудио-, видеоданных и управляющих сигналов по цифровым сетям. Существуют три основных конкурирующих между собой семейства протоколов, реализующих свой подход к построению мультимедийных мобильных и IP-сетей.

Исторически первый и наиболее распространённый в настоящее время протокол — H.323 был принят Международным союзом электросвязи (ITU) в 1996 г. H.323 представляет собой зонтичный протокол, который объединяет набор рекомендаций для мультимедийных приложений в цифровых вычислительных сетях, не обеспечивающих гарантированное качество обслуживания (Quality of Service, QoS). Рекомендации H.323 охватывают сервисы передачи аудио-, видео- и цифровых данных в сетях с коммутацией пакетов и специфицируют:

- управление полосой пропускания;
- стандарты для аудио- и видеокодеков;
- межсетевые конференции для разнородных сетей;
- поддержку многоточечных конференций — три и более участников;
- поддержку многоадресной передачи в многоточечной конференции;
- поддержку групповой адресации;
- кроссплатформенность.

Второй, не уступающий по распространённости протоколу H.323, протокол инициирования сеансов связи (Session Initiation Protocol, SIP), который впервые появился в марте 1999 г. и был разработан в рамках IETF. Протокол SIP описан в рекомендации RFC 2543. В отличие от H.323 это самостоятельный протокол прикладного уровня, который регламентирует установление, изменение и завершение мультимедийных сеансов связи с одним или несколькими участниками. SIP является клиент-серверным протоколом и работает на основе обмена последовательностью запросов-ответов, которые реализованы с помощью текстовых тегов, т.е. все SIP-заголовки передаются в виде ASCII-текста, что существенно упрощает его реализацию и использование в пользовательских приложениях.

И наконец, третий VoIP протокол — протокол контроля медиашлюзов (Media Gateway Control Protocol, MGCP), который появился в октябре 1999 г. в результате объединения двух протоколов — SGCP (Simple Gateway Control Protocol) и IPDC (Internet Protocol for Device Control), которые были разработаны компаниями Bellcore, Cisco Systems и Level 3. Протокол MGCP и родственные ему спецификации — SGCP, IPDC, MEGACO, H.248 основаны на жёсткой

иерархии, подразумевающей всего два функциональных компонента и полное отделение управления сигнализацией от медиапотоков. Управление сигнализацией осуществляется центральным управляющим устройством — контроллером сигнализаций, а медиапотоки обрабатываются шлюзами или абонентскими терминалами, например IP-телефонами. Функциональное назначение конечных исполнительных устройств — шлюзов (или абонентских терминалов) определяется набором понятных им команд, поступающих в простом текстовом формате от контроллера сигнализаций. Он же задаёт и ориентацию соединений между конечными устройствами на передачу голоса, факсимильных сообщений или цифровых данных.

Перечисленные протоколы позволяют осуществить управление передачей аудио-, видеоданных, устанавливать и управлять физическими и логическими соединениями, а также создавать каналы передачи данных. В качестве транспортного протокола для передачи мультмедийных данных используют протокол реального времени RTP (Real-time Transport Protocol), который был разработан Audio-Video Transport Working Group в IETF и впервые опубликован в 1996 г. Протокол RTP описан в спецификации RFC3550. Для передачи последовательности параметров речевого сигнала в PCPP были разработаны две дополнительные спецификации RFC3557 и RFC4060.

RTP содержит в своём заголовке данные, необходимые для восстановления речевого сигнала или видеоизображения в приёмном узле, а также данные о типе кодирования информации, например, тип речевого кодека. В заголовке данного протокола, в частности, передаётся временная метка и номер пакета. Эти параметры позволяют при минимальных задержках определить порядок и момент декодирования каждого пакета, а также обнаруживать потерянные пакеты. В качестве нижележащего протокола транспортного уровня, как правило, используется протокол UDP.

Для работы в компьютерных сетях необходимо выполнить адаптацию системы распознавания речи. Существенно облегчить эту задачу позволяет протокол управления медиаприложениями Media Resource Control Protocol (MRCP). Этот сетевой протокол был разработан для унифицированного управления по уже созданному каналу сигнализации мультимедийными ресурсами, такими как распознавание речи, синтез речи, верификация и идентификация диктора, а также цифровой диктофон. Стандарт MRCP был разработан Internet Engineering Task Force (IETF).

MRCP представляет собой одновременно инфраструктуру и протокол. Инфраструктура, показанная на **рис. 12**, определяет собой сетевые элементы и их взаимодействие друг с другом, а также с другими протоколами SIP и RTP. Так, при использовании УСРР клиенту необходимо выполнить следующие операции:

- найти удалённый сервер распознавания речи и запросить его возможности;
- установить физическое и логическое соединение, канал сигнализации с найденным сервером;
- осуществить передачу речевых аудиоданных и удалённое управление системой распознавания.

Для выполнения первой и второй процедуры MRCP использует протокол SIP, который позволяет клиенту найти необходимый сервер, используя SIP

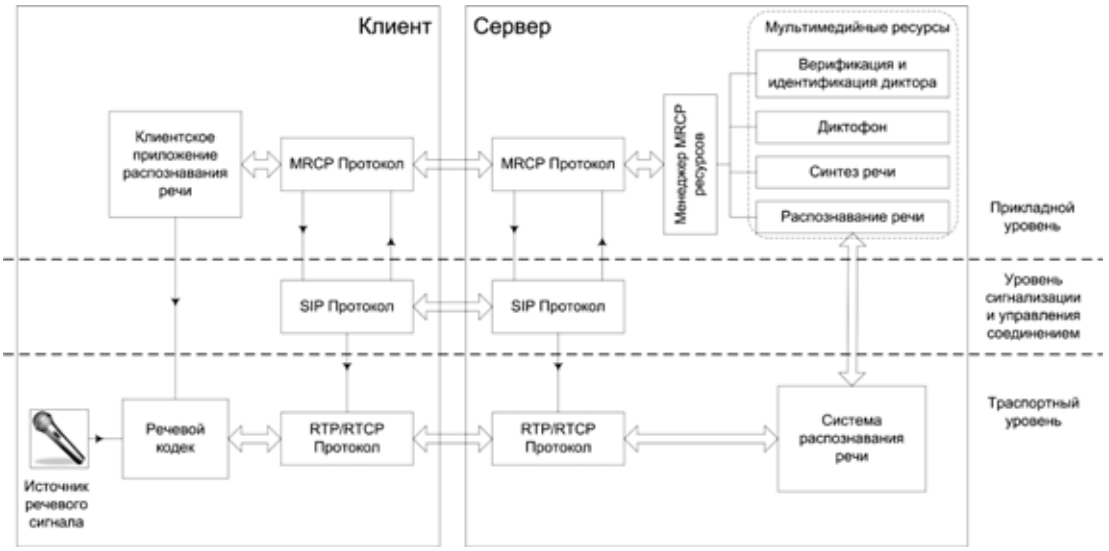

Рис. 12. Инфраструктура MRCP

Uniform Resource Indicator, а потом создать канал сигнализации с найденным сервером. Этот канал сигнализации необходим для обеспечения выполнения третьего пункта. По установленному каналу сигнализации MRCP осуществляет управляющее взаимодействие, т. е. обмен командами и ответами.

Заключение

В статье представлен краткий обзор принципов построения удалённой системы распознавания, основное преимущество которой — отсутствие ограничений на вычислительную мощность пользовательского компьютера. Также рассмотрены недостатки такой архитектуры, связанные прежде всего с повышением уровня шума и помехами, вносимыми цифровыми каналами связи. Проведён анализ методов, позволяющих снизить влияние этих недостатков на качество работы системы распознавания.

Кроме того, приведено краткое описание стандартов и протоколов VoIP, которые необходимы для реализации системы распознавания речи в среде цифровых вычислительных сетей.

Литература

1. *Rabiner L.R., Juang B.-H.,* Fundamentals of speech recognition. Prentice-Hall, Inc. 1993.
2. *Kato M., Sugiyama A., Serizawa M.,* Noise Suppression with High Speech Quality Based on Weighted Noise Estimation and MMSE STSA, IEICE Transaction, vol. E85-A, no. 7, July 2002.
3. *Boll S.F.,* Suppression of acoustic noise in speech using spectral subtraction, IEEE Transaction, ASSP-27, pp. 113–120, 1979.
4. *Lockwood P., Boudy J.,* Experiments with a non-linear spectral subtractor (NSS) hidden markov models and the projection, for robust speech recognition in car. Speech Communications, vol. 11, Issue 2–3, June 1992, pp. 215–228.

5. *Vaseghi S.V., Milner B.P.,* Noise compensation methods for hidden markov models speech recognition in adverse environments, IEEE Transaction on Speech and Audio Processing, vol. 5, no. 1, 1997, pp. 11–21.

6. *Lim J. S., Oppenheim A.V.,* All-pole modeling of degraded speech, IEEE Transaction, ASSP-26, 1978, pp. 197–210.

7. *Bernstein A. D., Shallom I. D.,* An hypothesized Wiener filtering approach to noisy speech recognition, // Proceedings of ICASSP-91., 1991, pp. 913–916. Toronto, Canada.

8. *Vaseghi S. V., Milner B. P.,* Noisy speech recognition based on HMM's, Wiener filters and re-evaluation of most likely candidates, // Proceedings of ICASSP-93, 1993, vol. 2, pp. 103–106. Minneapolis, MN, USA.

9. *Berouti M., Schwartz R., Makhoul J.,* Enhancement of speech corrupted by acoustic noise, // Proceedings of ICASSP-79, 1979, pp. 208–211. Washington DC, USA.

10. ETSI ES 202 050 v1.1.3. Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms. Technical Report ETSI ES 202 050, 2003, ETSI.

11. *Markel J., Gray A.,* Linear Prediction of Speech, Springer-Verlag, 1976. Русский перевод: *Маркел Дж.Д., Грэй А.Х.* Линейное предсказание речи, «Связь», Москва 1980.

12. *Davis S., Mermelstein P.,* Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transaction on Acoustics, Speech and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

13. *Oppenheim A., Schafer R.,* Digital Signal Processing. Prentice-Hall, Inc. 1975. Русский перевод: *Оппенгейм А.В., Шафер Р.В.,* Цифровая обработка сигналов, «Связь», Москва 1979.

14. *Рабинер Л.Р., Шафер Р.В.,* Цифровая обработка речевых сигналов. М.: Радио и связь, 1981.

15. *Oppenheim A.V., Johnson D.H.,* Discrete representation of signals. // Proceedings of the IEEE, vol. 60, no. 6, pp. 681–691. 1972.

16. *Wölfen M., McDonough J.,* Minimum variance distortionless response spectral estimation, IEEE Signal Processing Magazine, vol. 22, no. 5, pp. 117–126. 2005.

17. *Furui S.,* Cepstral analysis technique for automatic speaker verification, IEEE Transaction on ASSP-29, pp. 254–272, 1981.

18. *Furui S.,* Speaker-independent isolated word recognition using dynamic features of speech spectrum, IEEE Transaction on ASSP-34, pp. 52–59, 1986.

19. *Young S., Jansen J., Odell J., Ollason D.,* Woodland P., HTK — Hidden Markov Toolkit, Entropic Cambridge Research Laboratory, 1995.

20. *Acero, A., Crespo, C., De la Torre, C., Torrecilla, J.C,* Robust HMM-based endpoint detector. // Proceedings of the EuroSpeech'93, Berlin, 21-23 September 1993. vol. 3. pp. 1551–1554, 1993.

21. *Rabiner L.R.* A tutorial on hidden Markov models and selected application in speech recognition, Proceedings of the IEEE, vol. 77, 2, 1989, pp. 257–286. Русский перевод: *Рабинер Л.Р.* Скрытые марковские модели и их применение в избранных приложениях при распознавании речи: Обзор ТИИЭР 1989 Т. 77, № 2. С. 86–120.

22. *Varga I., Aalburg S., Andrassy B., Astrov S., Bauer J.G., Beaugeant Ch., Geissler Ch., Hode, H.* ASR in mobile phones — An industrial approach. IEEE Transaction on Speech and Audio Processing, vol. 10, no. 8, 2002, pp. 562–569.

23. *Deligne S., Dharanipragada S., Gopinath R., Maison B., Olsen P., Printz H.* A robust high accuracy speech recognition system for mobile applications. IEEE Transaction on Speech and Audio Processing, vol. 10, no. 8, 2002, pp. 551–561.

24. *Haeb-Umbach R.* Robust speech recognition for wireless networks and mobile telephony, // Proceedings of the Eurospeech'97, Rhodes, Greece, 1997.

25. *Goldberg R., Riek L.* A Practical Handbook of Speech Coders. CRC Press, Boca Raton, FL, 2000.

26. ETSI EN 300 961 GSM Full Rate Speech Transcoding (GSM 06.10). Technical Report ETSI EN 300 961, 1995, ETSI.

27. ETSI EN 300 726 Enhanced Full Rate (EFR) Speech Transcoding (GSM 06.60). Technical Report ETSI EN 300 726, 1999, ETSI.

28. ETSI EN 301 704 Adaptive Multi-Rate (AMR) Speech Transcoding (GSM 06.90). Technical Report ETSI EN 301 704, 1998, ETSI.

29. *Euler S., Zinke J.* The influence of speech coding algorithms on automatic speech recognition. // Proceedings of ICASSP-94, 1994, vol. 1, pp. 621–624, Adelaide, Australia.

30. *Sukkar R.A., Chengalvarayan R. and Jacob J.J.* Unified speech recognition for the landline and wireless environments. // Proceedings of ICASSP-02, 2002, pp. 293-296, Orlando, Florida, USA.

31. *Lilly B.T., Paliwal K.K.* Effect of speech coders on speech recognition performance. // Proceeding of ICSLP, 1996, pp. 2344-2347, Philadelphia, PA, USA.

32. *Pelaez-Moreno C., Gallardo-Antolin A., Diaz-de-Maria F.* Recognizing voice over IP: A robust front-ends for speech recognition on the World Wide Web. IEEE Transaction on Multimedia, vol. 3, no. 2, 2001, pp. 209–218.

33. *Kim, H.K., Rose R.C.,* Speech recognition over mobile Networks. // Automatic speech recognition on mobile devices and over communication networks. Springer, 2008.

34. *Kim, H. K. and Cox, R. V. A.* Bitstream-based front-end for wireless speech recognition on IS-136 communication system. IEEE Transaction on Speech and Audio Processing, vol. 9, no. 5, pp. 558–568, 2001.

35. *Tan Z.-H., Dalsgaard, P., Lindberg B.* Exploiting temporal correlation of speech for error robust and bandwidth flexible distributed speech recognition. IEEE Transactions on Audio Speech and Language Processing, vol. 15, no. 4, pp. 1391–1403, 2007.

36. *Huerta J.M., Stern R.M.* Speech recognition from GSM codec parameters. // Proceedings of ICSLP, 1998, pp. 1463–1466.

37. *Gallardo-Antolin A., Diaz-de-Maria F., Valverde-Albacete F.* Recognition from GSM digital speech. // Proceedings of ICSLP, 1998, pp. 1443–1446.

38. *Gallardo-Antolin A., Pelaez-Moreno C., Diaz-de-Maria F.* Recognizing GSM digital speech. IEEE Transaction on Speech and Audio Processing, vol. 13, no. 6, 2005, pp. 1186–1205.

39. *Pearce, D.* Enabling New Speech Driven Services for Mobile Devices: An Overview of the ETSI Standards Activities for Distributed Speech Recognition Front-ends. Applied Voice Input/Output Society Conference (AVIO2000), San Jose, CA, May 2000.

40. ETSI ES 201 108 v1.1.3. Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms. Technical Report ETSI ES 201 108, 2003, ETSI.

41. ETSI ES 202 050 v1.1.3. Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms. Technical Report ETSI ES 202 050, 2003, ETSI.

42. ETSI ES 202 211 v1.1.1. Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm. Technical Report ETSI ES 202 211, 2001, ETSI.

43. ETSI ES 202 212 v1.1.1. Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction. Technical Report ETSI ES 202 212, 2003, ETSI.

44. ETSI TS 126 243 — UMTS; ANSI-C code for the fixed-point distributed speech recognition extended advanced front-end. Technical Report ETSI TS 126 243, 3GPP 2004, ETSI.

45. *Bossert, M.* Channel Coding for telecommunication. John Wiley & Sons. 2000.

46. *Haeb-Umbach R and Ion V.* Error Concealment. // In Automatic Speech recognition on mobile devices and over communication network. Springer 2008.

47. *Kim, H. K.,* Speech recognition over IP networks. // In Automatic Speech recognition on mobile devices and over communication network. Springer 2008.

48. *Bernard A., Alwan A.* Low-bitrate distributed speech recognition for packet-based and wireless communication. IEEE Transactions on Speech and Audio Processing, vol. 10, no.8, pp. 570–579, 2002.

49. *Tan Z.-H., Dalsgaard, P., Lindberg B.* Automatic speech recognition over error-prone wireless network. Speech Communication, vol. 47, no. 1–2, pp. 220–242, 2005.

50. *Boulis C., Ostendorf M., Riskin E. A., Otterson, S.* Graceful degradation of speech recognition performance over packet-erasure networks. IEEE Transaction on Speech Audio Processing, vol. 10, no. 8, pp. 580–590, 2002.

51. *Viterbi A.,* Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm, IEEE Transaction, 1967, IT-13, pp.260–269.

52. *Potamianos A., Weerackody V.* Soft-feature decoding for speech recognition over wireless channels. // Proceedings of ICASSP-01, 2001, vol. 1. Salt Lake City, Utah, USA.

53. *Pienado A.M., Gomez A.M., Sanchez V.* Packet Loss Concealment Based on VQ Replicas and MMSE Estimation Applied to Distributed Speech Recognition. // Proceedings of ICASSP-05, 2005, vol. 1, pp. 329–330, Philadelphia, PA, USA.

54. *Weerackody V., Reichl W., Potamianos A.* An Error-Protected Speech Recognition System for Wireless Communications. IEEE Transaction on Wireless Communications, vol. 1, No. 2, 2002, pp. 282–291.

55. *Delaney B.* Increased robustness against bit errors for distributed speech recognition in wireless environments. // Proceedings of ICASSP-05, 2005, vol. 1. pp. 313–316, Philadelphia, PA, USA.

56. *James A., Milner B.* An analysis of interleavers for robust speech recognition in burst-like packet loss. // Proceedings of ICASSP-04, 2004, vol. 1, pp. 853–856. Montreal, Canada.

57. *Milner B., Semnani S.* Robust speech recognition over IP networks. // Proceedings of ICASSP-2000, 2000, vol. 3, pp.1791–1794, Istanbul, Turkey.

58. *James A., Gomez A., Milner B.,* A comparison of packet loss compensation methods and interleaving for speech recognition in burst-like packet loss. // Proceedings of ICASSP-04, 2004, Jeju Island, Korea.

59. *Pienado A.M., Segura J.C.* Speech recognition over digital channels. John Wiley & Sons. 2006.

60. *Fingscheidt T., Aalburg S., Stan S., Beaugeant C.* Network-based vs. distributes speech recognition in adaptive multi-rate wireless systems. // Proceedings of ICASSP-02, 2002, Denver, USA.

61. *Ion V., Haeb-Umbach R.* A unified probabilistic approach to error concealment for distributed speech recognition. // Proceedings of Eurospeech'2005, Lisbon, Portugal.

62. *Kelleher H., Pearce D., Ealy D., Mauuary L.* Speech recognition performance comparison between DSR and AMR transcoded speech. // Proceedings of ICSLP'2002. Denver, USA.

63. *Kiss I.* A comparison of distributed and network speech recognition for mobile communication systems. // Proceedings of ICSLP'2000. Beijing, China.

Маковкин Константин Александрович —
окончил Московский государственный технический университет им. Н.Э. Баумана в 1990 г. С 1990 года сотрудник Вычислительного центра им. А. А. Дородницына РАН. Область интересов: разработка систем автоматического распознавания речи; цифровая обработка сигналов; скрытые марковские модели; модели нейронных сетей; VoIP протоколы
E-mail: k.makovkin@gmail.com