Доклады БГУИР
2006 *январь–март* *№ 1 (13)*

КРАТКИЕ СООБЩЕНИЯ

УДК 621.391

ТЕНДЕНЦИИ РАЗВИТИЯ РЕЧЕВЫХ РАСПОЗНАЮЩИХ СИСТЕМ В МУЛЬТИМЕДИЙНОЙ ТЕЛЕФОНИИ

А.С. РЫЛОВ

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

Поступила в редакцию 30 января 2006

В статье обсуждаются два основных направления для создания речевых распознающих систем в мультимедийной телефонии. Кроме того, рассмотрены основные требования к таким распознавателям и ограничения, накладываемые на них телефонным каналом.

Ключевые слова: распознавание речи, телекоммуникационные системы, признаковое пространство, мультимедийная телефония.

Введение

Компьютерная телефония является одной из основных составляющих современных телекоммуникационных систем. Эта технология интегрирует компьютер в телефонную сеть, при этом пользователь получает доступ к широкому спектру телекоммуникационных средств, таких как модем, факс, телефон, голосовая и электронная почта, доступ к сетям и удаленным базам данных [1]. Дальнейший прогресс современных телекоммуникационных систем наряду с другими направлениями предполагает также развитие мультимедийной телефонии и, прежде всего, речевых распознающих систем. Сегодня уже не стоит вопрос быть или не быть такого рода интеграции. Более того, она близка к завершению в США, Японии, некоторых странах Европы. Встроенные системы распознавания речи уже давно используются в мобильных телефонах в качестве номеронабирателя. Широко используется сочетание систем верификации личности по парольной фразе и распознавателей речевых команд, применяемых в качестве оконечных устройств в банковских телекоммуникационных системах (home banking), позволяющих клиентам банка, не выходя из дома, осуществлять кредитные операции, а также производить оплату за коммунальные услуги и т.п. Эти и другие известные примеры подтверждают, что введение речевых распознающих систем в качестве оконечных телекоммуникационных устройств придает системам связи качественно новые свойства.

Компактность признаковых пространств

Встроенные в телекоммуникационные системы распознающие устройства должны быть не слишком сложными и дорогими и при этом обладать высокой надежностью распознавания. Это возможно только в том случае, если самые начальные уровни иерархической системы распознавания [2] обеспечивают формирование компактных пространств признаков, которые заключают в себе максимально возможное количество информации при минимально возможной их размерности и перекрываемости.

На рис. 1 показана межклассовая перекрываемость классов образов A, B, C, D, которая должна быть минимальной за счет внутриклассовой компактности признаковых пространств.

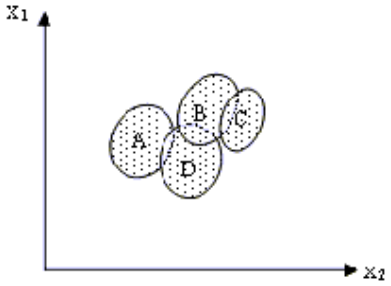

Рис. 1. Двухмерные пространства признаков классов образов A, B, C, D

Таким образом, преобразование исходных данных в информативные признаки обеспечивает успешное распознавание даже при использовании относительно простых решающих правил, тогда как в "плохом" пространстве невозможно достичь высокой надежности даже с помощью самых изощренных алгоритмов" [3]. Это означает, что при достижении определенной степени внутриклассовой компактности признаковых пространств, автоматически обеспечивается высокая степень межклассовых различий пространств признаков. В результате этого появляется возможность создавать более простые устройства распознавания, пригодные для использования их в телекоммуникационных системах в качестве встроенных модулей в оконечных устройствах объектов связи.

Степень разделимости классов образов или информативность параметров x_1, x_2,…, x_n может быть оценена с помощью информационной меры-дивергенции [4]. Последняя в случае нормального распределения функции плотности вероятности сравниваемых параметров классов образов W_i и W_j и равенства их ковариационных матриц

$$V_i = V_j = V \tag{1}$$

превращается в меру Махаланобиса

$$d_m = (M_i - M_j)^T V^{-1} (M_i - M_j)\,. \tag{2}$$

Отличие заключается лишь в том, что в (2) берется разность между двумя средними M_i и M_j параметров. Поэтому (2) может быть одновременно мерой близости между двумя ВК-пространствами признаков [5] классов образов и показателем эффективности тех или иных разновидностей векторов параметров, являющихся центроидами кодовых книг векторного квантования. Эффективность векторов-параметров может оцениваться по значению обобщенной ошибки (EER) (рис. 2) [6], получаемой в результате их тестирования при использовании какой-либо речевой базы данных.

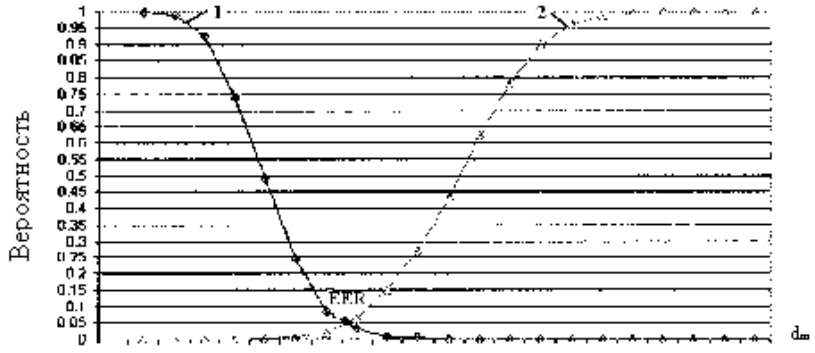

Рис. 2. Распределения вероятностей несовершенных ошибок 1-го и 2-го рода

Значение EER является точкой пересечения распределений вероятностей несовершения ошибок 1-го и 2-го рода в зависимости от значения меры (2) [4]. Кривая 1 характеризует внутриклассовую вероятность, а кривая 2 — межклассовую. При EER=0 перекрытие между классами образов отсутствует. Эти кривые рассчитываются с помощью гистограмм распределения внутриклассовых — 1 и межклассовых — 2 мер близости (2), представленных на рис. 3.

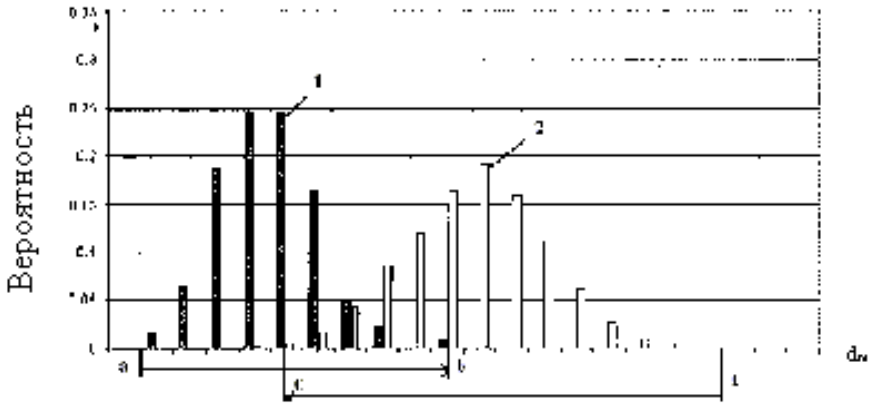

Рис. 3. Гистограммы распределений внутриклассовых и межклассовых мер близости

Это осуществляется путем поочередного вычитания текущих значений вероятностей на рис. 3 из предыдущих, начиная с точек *a* и *d* соответственно, для которых вероятность берется равной 1. Убывание значений вероятностей на рис. 2 идет в направлениях, указанных стрелками на рис. 3.

В таблице показаны результаты расчета гистограмм (первые три строчки) и распределений вероятностей несовершения ошибок 1-го и 2-го рода (последние две строчки). При этом EER=0,048. Она характеризует в данном случае эффективность Δ-сепстральных параметров для текстозависимой системы верификации личности по речевому сигналу [6].

Результаты расчета гистограмм, EER и распределений вероятностей несовершения ошибок 1-го и 2-го рода

d_m	интервалы	6,404	7,706	9,007	10,31	11,61	12,91	14,21	15,52	16,82	18,12	19,42	20,72	22,02	23,33	24,63	25,93	27,23	28,53	29,83
Вероятность	внутриклас.	0,014	0,064	0,186	0,245	0,245	0,164	0,05	0,023	0	0,009	0	0	0	0	0	0	0	0	0
	межклас.	0	0	3E-04	0,002	0,004	0,015	0,043	0,086	0,121	0,163	0,191	0,159	0,113	0,063	0,026	0,01	0,003	3E-04	3E-04
	н/сов.ош. 1-го рода	0	0	0	3E-04	0,002	0,006	0,021	0,064	0,149	0,27	0,434	0,625	0,784	0,897	0,96	0,986	0,996	0,999	0,999
	н/сов.ош. 2-го рода	0,995	0,986	0,923	0,736	0,491	0,245	0,082	0,032	0,009	0,009	0	0	0	0	0	0	0	0	0

Таким образом, по величине обобщенной ошибки EER, тестируя различные параметры, можно оценить информативность каждого параметра, входящего в состав вектора параметров.

Отметим, что все выше изложенное относится как к системам распознавания речи, так и к системам распознавания личности по речевому сигналу, которые могут быть использованы для предъявления речевого биометрического паспорта в IP-телефонии. При отсутствии возможности применять сложные алгоритмы классификаторов в устройствах распознавания речи, предназначенных для мультимедийной телефонии, последние должны иметь ограничения на объем словаря и на слитность произношения.

Передача речевых сообщений по сверхузкополосному каналу

Важным аспектом мультимедийной телефонии является передача речевых сообщений при минимально возможном количестве информации (десятки, единицы бит на секунду речи). К середине 80-х – началу 90-х годов прошлого столетия было установлено [7], что компрессия

речевых сигналов после 1200–600 бит/с для передачи по сверхузкополосному каналу может осуществляться только на основе распознавания каких-либо единиц речи (фонем, слогов, слов). Основанием для этого вывода послужило то, что в результате проведенных экспериментов была отвергнута идея непосредственного преобразования параметров речевого сигнала в последовательность единиц фонемной размерности. Стало ясно, что дальнейшее снижение объема информации о речевом сигнале может быть осуществлено только методом лингвистической компрессии, т.е. путем распознавания речеязыковых единиц (фонем, слогов, слов). Лингвистическая компрессия предполагает замену анализатора речи (параметризатора) в передающей части вокодерной системы на распознаватель определенных единиц речи (фонем, слогов, слов и даже словосочетаний). На приемной же стороне вместо параметрического синтезатора речи может использоваться устройство для воспроизведения заданного набора (словаря) предварительно записанных единиц речи. На рис. 4 для сравнения представлены структурные схемы: а) традиционного вокодера с параметрической компрессией речевого сигнала; б) вокодера с лингвистической компрессией речевого сигнала на основе распознавания определенных единиц речи.

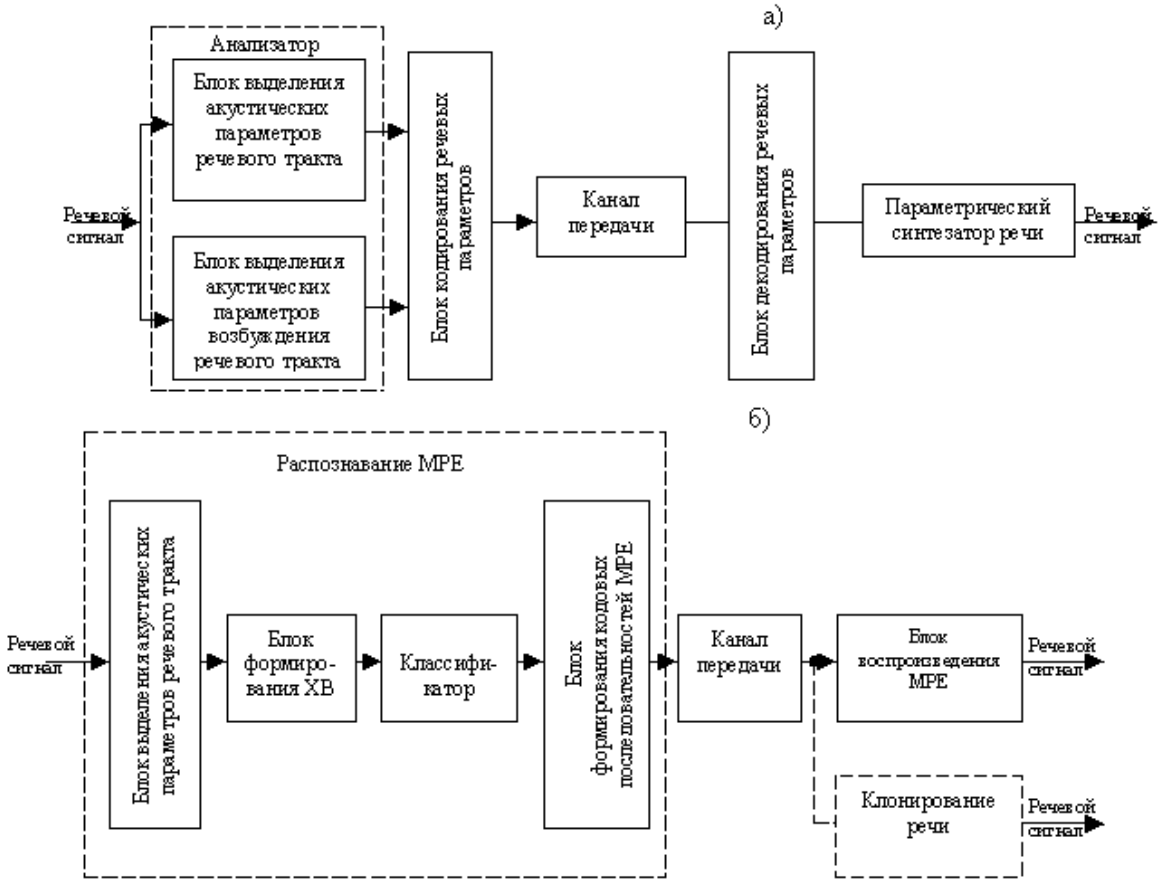

Рис. 4. Структурные схемы: а) вокодер с параметрической компрессией речевого сигнала; б) вокодер с лингвистической компрессией речевого сигнала

В лингвистическом вокодере (на рис. 4,*б*) после блока выделения акустических параметров речевого тракта стоит блок формирования характеристических векторов (ХВ), обеспечивающий компактность признаковых пространств с целью упрощения алгоритмов классификации минимальных речевых единиц (МРЕ) в классификаторе. При этом в канал связи будут поступать коды, каждый из которых соответствует определенной единице из заданного словаря. Поэтому если использовать в качестве единиц речи слова, то при определенном ограничении объема словаря и раздельном их произношении можно снизить информационный поток

сообщения в канале до десятков бит/с. Если в качестве МРЕ использовать более мелкие единицы речи (фонемы, дифоны, трифоны, слоги), то в этом случае для восстановления речи на приемном конце можно применить алгоритмы волнового синтеза или клонирования речи [8]. Однако сложность алгоритмов работы классификатора резко возрастает, но значительно снижаются ограничения на форму передаваемой речи, т.е. на слитность и объем словаря. При этом количество информации на одну секунду передаваемой речи может быть снижено до 300 бит/с. За счет такого существенного снижения количества информации на единицу времени можно применять различные методы помехоустойчивого кодирования с тем, чтобы добиться высокой надежности передаваемых речевых сообщений в зашумленных каналах связи.

Заключение

В заключение можно сделать следующие выводы:

– речевые распознающие системы в мультимедийной телефонии обладают высокой надежностью распознавания при наличии ограничений на объем словаря и слитность произношения речи;

– сложность алгоритмов работы классификаторов распознающих систем в мультимедийной телефонии должна компенсироваться применением эффективных методов формирования компактных признаковых пространств на начальных уровнях иерархической системы распознавания;

– распознавание речевых образов в мультимедийной телефонии может быть использовано как для создания речевого интерфейса, так и для создания специальных сверхузкополосных систем связи.

DEVELOPING TENDENCY OF SPEECH RECOGNITION SYSTEMS IN MULTIMEDIA TELEPHONY

A.S. RYLOV

Abstract

Two principal directions for creation of speech recognition systems in multimedia telephony are discussed. Besides the principal demands and putted on telephone channel limitations for these recognizers are considered.

Литература

1. *Чучупал* В.Я., *Маковкин* К.А. // Современные речевые технологии: Сб. трудов IX сессии Российского акустического общества, М., 26–28 января 1999 г. М., 1999. С. 81–84.
2. *Рылов* А.С. // Весці НАН Беларусі. Сер. фіз.-тэхн. навук. 2000. № 2. С. 100–114.
3. *Сорокин* В.Н. // Современные речевые технологии: Сб. трудов IX сессии Российского акустического общества. Москва, 26–28 января 1999 г. М., 1999. С. 50–55.
4. *Рылов* А.С., *Чижденко* В.А., *Левковская* Т.В. // Докл. БГУИР. 2004. № 6. С. 39-44.
5. *Рылов* А.С. // Докл. НАН Беларусі. 2004. Т. 48, № 4. С. 38–41.
6. *Рылов* А.С., *Чижденко* В.А. Заявка a20030890 на патент с приоритетом от 23.09.2003. Опубл. БИ, № 3, 2004.
7. *Белявский* В.М., *Зеленый* А.И. // Автоматическое распознавание слуховых образов (АРСО-17): Тез. докл. 17-й Всесоюз. школы-семинара. Ижевск, 1992. С. 25-27.
8. *Лобанов* Б.М. // Новости искусственного интеллекта. 2002. № 5 (55). С. 35–39.