Речевые технологии

2/2008

Содержание

ЗАМЕТКИ — ОБЗОРЫ

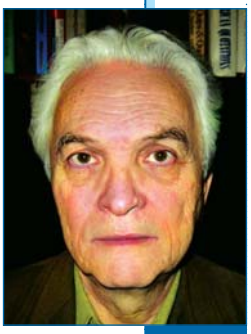
ИСТОРИЯ

Редакция:
Редактор *Ольга Подколзина*
Корректор *Татьяна Денисьева*
Дизайн и вёрстка *Анна Ладанюк, Максим Буланов*

Адрес редакции: 109341, Москва, ул. Люблинская, д. 157, корп. 2.
Тел.: 8 901 510-30-65

Подписано в печать 15.01.2009. Формат 60×90⅛. Бумага офсетная. Печать офсетная. Печ. л. 6. Заказ № 0628. Издательский дом «Народное образование».
Отпечатано в типографии НИИ школьных технологий. 143500, г. Истра-2, ул. Заводская, д. 2А. Тел.: 8 901 513-97-64, (495) 792-59-62.

Автокорреляционный метод выделения основного тона речи. Пятьдесят лет спустя



С.П. Баронин,
кандидат технических наук

Проблема выделения основного тона речи рассматривается с позиций теории статистических решений. Основываясь на модели вокализованного речевого сигнала в виде периодического процесса с неизвестными параметрами, искажённого шумом, и принимая в качестве модели сигнала основного тона нестационарный Марковский процесс, предложена разновидность автокорреляционного выделителя основного тона. Моделирование алгоритма и испытания на тестовых сигналах подтвердили высокую точность измерения периодов основного тона. Приводятся также данные о помехоустойчивости описанного алгоритма.

Немного истории

Пятьдесят лет назад учёный с мировым именем Андрей Андреевич Пирогов получил авторское свидетельство на изобретение устройства для определения частоты основного тона (ОТ) по методу автокорреляционного анализа речевого сигнала [1]. В то же время в руководимой А.А. Пироговым лаборатории начаты работы по теоретическому и экспериментальному исследованию автокорреляционного метода выделения ОТ речи [2]. В 1959 году создан первый макет устройства. Испытания макета подтвердили правильность основной идеи метода, хотя аппаратурная реализация на элементной базе того времени представляла большие трудности. Интегральных схем ещё не было, транзисторы с весьма посредственными характеристиками только появлялись и были дефицитом, так что в первом макете использовались лампы, диоды и множество катушек. Линия задержки содержала 150 звеньев LC, параметры которых для исключения отражений приходилось подбирать с высокой точностью. Вместо умножителей использовались ключевые схемы на диодах, так что вместо автокорреляционной функции использовалась функция взаимной корреляции речевого сигнала (РС) и клиппированного РС. В 1960–1967 годы были созданы макеты на транзисторах с использованием элементов цифровой техники. Однако широкое использование автокорреляционных выделителей основного тона (ВОТ) в составе вокодеров началось с появлением БИС и микропроцессоров [3–5].

Первые макеты автокорреляционных ВОТ работали удовлетворительно, правда, при подавлении в РС частот ниже 300 Гц (коммутируемый телефонный канал) количество ошибок резко возрастало. Попытки улучшить работу путём различных усовершенствований схем и алгоритмов не давали желаемых результатов. Выявилась необходимость разработки общего подхода к проблеме выделения ОТ с позиций статистической теории оценивания параметров случайных сигналов [6–10].

Статистический подход к проблеме выделения ОТ

Речевой сигнал представляет собой нестационарный случайный процесс с неизвестными параметрами, один из которых — основной тон. Обычно для удобства анализа РС разбивается на квазистационарные сегменты длительностью 10–45 мс и на каждом сегменте параметры сигнала считаются постоянными. РС на вокализованном сегменте можно представить в виде суммы гармоник частоты основного тона с априори неизвестными амплитудами и фазами. Отклонения реального РС от такой математической модели, обусловленные нестационарностью РС в пределах сегмента и внешними помехами, будем считать шумом. Задача анализа РС сводится, таким образом, к статистической задаче оценки неизвестных параметров случайного процесса, искажённого шумом. При определённых допущениях [6, 8] оптимальным алгоритмом оценки периода ОТ на рассматриваемом сегменте является вычисление автокорреляционной функции (АКФ) РС и определение аргумента, при котором АКФ максимальна.

Достоверное определение значения ОТ по одному сегменту иногда оказывается невозможным. Существенное повышение точности может быть достигнуто при учёте зависимостей значений периодов ОТ на соседних сегментах. Обычно периоды ОТ на соседних сегментах близки, хотя бывают и исключения. Оптимальный алгоритм оценки периода ОТ на данном сегменте должен учитывать оценки периодов ОТ на соседних сегментах и вероятности изменений периодов ОТ.

Простейшей моделью движения ОТ является Марковский процесс, для такой модели алгоритм обработки значений АКФ оказывается достаточно простым и сводится к фильтрации значений АКФ линейным фильтром с импульсной реакцией, определяемой функцией вероятностей изменений периодов ОТ на соседних сегментах. Результаты фильтрации на соседних сегментах используются в качестве априорной информации при определении периода ОТ на данном сегменте [6–10].

Марковский процесс первого порядка (случайные блуждания типа броуновского движения) представляет весьма грубую модель изменений периодов ОТ реального РС. Тем не менее даже такая простая модель позволяет существенно повысить точность измерений. Дальнейшее повышение точности может быть достигнуто увеличением порядка модели. Модель второго порядка, учитывающая не только предыдущее значение ОТ, но и тенденцию изменений ОТ, больше соответствует характеристикам РС. Изменения периодов ОТ на соседних обычно невелики, хотя встречаются и значительные изменения, так что вероятности переходов плохо описываются гауссовским законом, более точное описание даёт сумма двух гауссовских распределений с разными дисперсиями [7, 8]. Вероятности больших изменений увеличиваются на границах вокализованных участков речи

и при изменениях характера звука, так что более совершенная модель движения ОТ представляется нестационарным процессом с изменяющейся дисперсией функции вероятностей переходов. Учёт этих особенностей позволяет повысить точность работы ВОТ.

Критерии периодичности

Общим определением периодичности процесса является повторяемость формы сигнала. Минимальный временной сдвиг, при котором достигается повторяемость формы, называется периодом. Посторонние шумы и нестационарность РС исключают возможность точного повторения формы через период ОТ. В связи с этим встаёт вопрос о критерии сравнения исходного и сдвинутого во времени фрагментов РС при определении периодичности. Для шумов с гауссовским распределением следует использовать среднеквадратический критерий, для этого критерия оптимальна автокорреляционная обработка. Однако внешние шумы (комнатные, шумы телефонной линии) обычно отличаются от гауссовских. Для выяснения свойств внутренних шумов, учитывающих отклонение реального РС от модели периодического процесса, проведён следующий эксперимент.

На вокализованных участках вычислялись абсолютные значения разностей между отсчётами РС, разнесёнными на период ОТ. Частота дискретизации 16 кГц получена интерполированием отсчётов РС [11] с частотой дискретизации 8 кГц. Для исключения влияния вариаций уровня на результаты измерений сигналы нормировались таким образом, чтобы средние уровни сигналов на сравниваемых фрагментах были одинаковы. На рис. 1 чёрным цветом представлена функция распределения PR абсолютных значений разностей отсчётов, отстоящих на период ОТ. По оси абсцисс отложены величины DS разностей отсчётов, нормированные относительно среднего абсолютного значения сигнала на периоде ОТ.

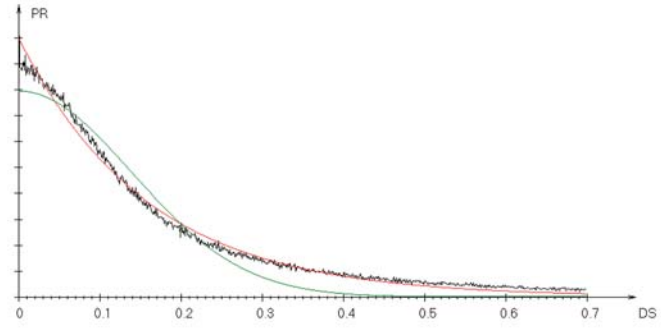


Рис. 1. Распределение разностей значений РС при сдвиге на период ОТ

Полученное распределение хорошо аппроксимируется функцией вида $pr(ds) = \alpha \times \exp(-\beta \times |ds|)$, где α, β — постоянные коэффициенты (красная кривая). Функция гауссовского распределения $pr(ds) = \lambda \times \exp(-\mu \times ds^2)$ (зелёная кривая) хуже представляет это распределение (ошибка аппроксимации при оптимальных значениях параметров λ, μ больше в 5,8 раза). Из этого можно заключить, что при оценке периода ОТ лучше пользоваться не критерием минимизации среднего квадрата отклонений (что эквивалентно максимизации АКФ), а критерием минимизации средней величины абсолютных значений отклонений. Соответствующий вариант разностной (сдвиговой) функции $S(t, \tau)$ [12, 13], по максимуму которой определяется период ОТ, может быть записан в виде

$$S(t,\tau) = \mathrm{am} \Big/ \left(\sum_{i=1}^{\tau} | x(t+i) - gain \times x(t+\tau+i) | + \delta \times am \right), \quad (1)$$ где

$$\mathrm{am} = \sum_{i=1}^{\tau} | x(t+i) |, \quad \mathrm{gain} = \mathrm{am} \Big/ \sum_{i=1}^{\tau} | x(t+\tau+i) |.$$

Параметр δ (в наших экспериментах δ=0,0625) позволяет регулировать соотношение между максимумами функции $S(t,\tau)$.

На *рис. 2, 5* показан вид разностных (сдвиговых) функций (РСФ) S(t,τ) и АКФ R(t,τ). Функции вычислены для одного и того же сегмента сигнала x(t), отсчёты по t и τ (tau) берутся через 1/16 кГц = 0,0625 мс. Пики у РСФ более острые, чем у АКФ, поэтому увеличивать интервал между отсчётами до часто используемого значения 1/8 кГц = 0,125 мс нежелательно. Положение пиков на оси τ у этих функций отличается обычно незначительно, но соотношение амплитуд пиков разное. Пик, соответствующий периоду ОТ, у РСФ чаще всего более выражен. К достоинствам РСФ можно отнести также меньшие вычислительные затраты — многочисленные умножения, производимые при вычислении АКФ, заменяются более простыми операциями сложения и отбрасывания знака у результата сложения.

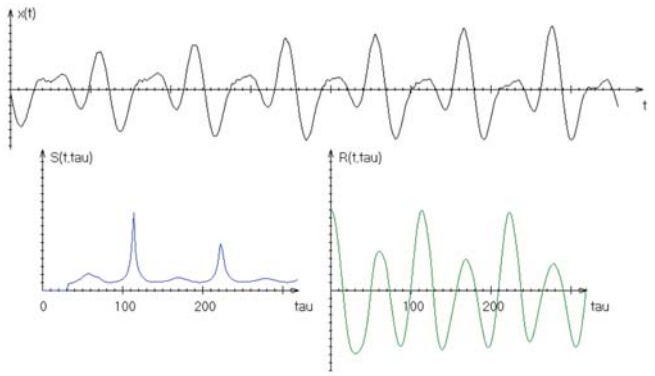


Рис. 2. Пример функции S (t,tau) и R (t,tau) сигналах x(t)

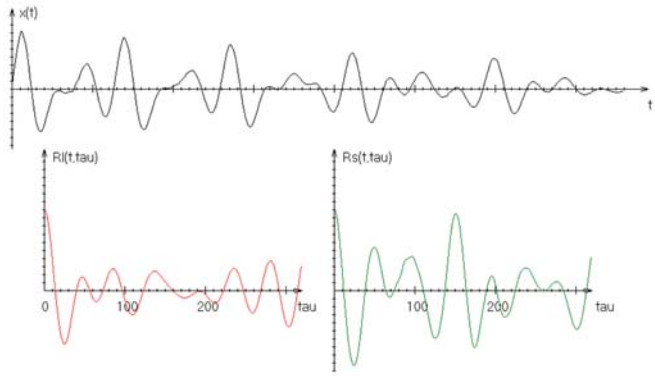


Рис. 3. Пример функции RI (t,tau) и Rs (t,tau) при быстром увеличении периодов ОТ сигнала x(t)

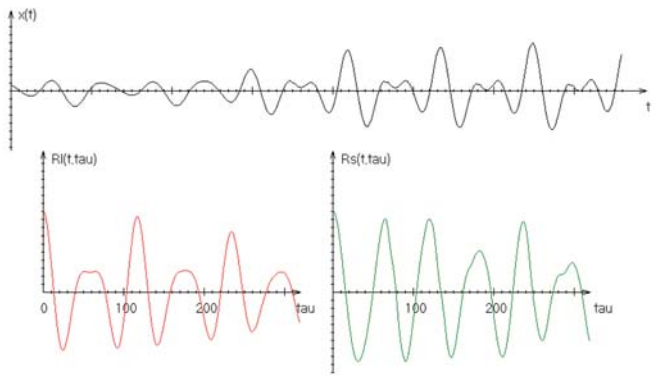


Рис. 4. Пример функции RI (t,tau) и Rs (t,tau) при резком изменении формы сигнала x(t)

Время усреднения

При вычислении АКФ время усреднения произведений x(t)×x(t+τ) обычно постоянно и составляет 10–45 мс. В данной работе при вычислении АКФ и РСФ время усреднения переменное и равно предполагаемому периоду ОТ, для которого вычисляется значение функции. Благодаря этому появляется возможность измерять не среднее значение периода на сегменте, а значения каждого периода ОТ. На нестационарных участках при быстрых изменениях периода ОТ разница этих двух методов существенна. На рис. 3 приведён пример АКФ Rl(t,τ) (время усреднения 45 мс) и АКФ Rs(t,τ) (время усреднения для каждого отсчёта функции равно предполагаемому значению периода ОТ) для РС x(t) с быстрым изменением ОТ. У функции Rs(t,τ) пик, соответствующий периоду ОТ, явно выражен.

При большом времени усреднения значения АКФ более устойчивы. Если период ОТ за время усреднения изменяется незначительно, а форма речевой волны резко изменяется (рис. 4), то пик функции Rl(t,τ) при τ равном периоду ОТ более выражен, чем пик функции Rs(t,τ). Тем не менее, представляется целесообразным на первом этапе обработки вычислять АКФ для каждого периода ОТ (минимальное время усреднения), а функции повышения достоверности за счёт обработки данных на большом интервале времени выполнять на втором этапе путём нелинейной фильтрации замеров АКФ или РСФ.

Учёт периодичности АКФ и РСФ

Для периодического процесса АКФ и РСФ имеют пики при значениях τ, кратных периоду ОТ. В реальном РС вследствие нестационарности амплитуды пиков по мере увеличения τ обычно уменьшаются, однако, бывают исключения и пик, например, при задержке на 2×τ может быть больше пика при задержке на время τ, равное периоду ОТ. Для исключения таких ошибок иногда значения АКФ умножают на некоторую монотонно уменьшающуюся функцию от τ. Это даёт некоторый эффект, хотя вероятность ошибочных выборов пиков на задержках, меньших периода ОТ, увеличивается.

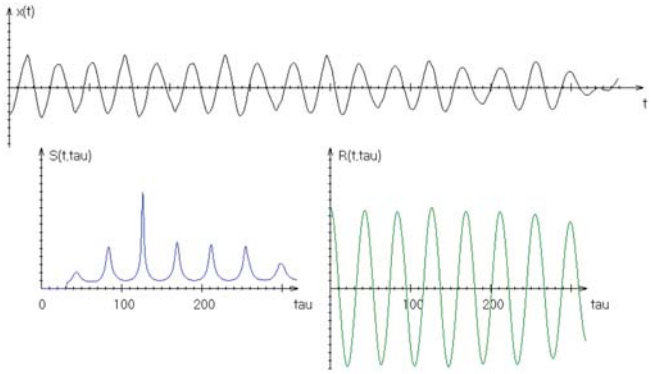


Рис. 5. Пример функции S (t,tau) и R (t,tau) сигнала x(t) с подавлением частот ниже 300 Гц

Из теории статистического оценивания следует, что при измерении периода ОТ надо учитывать не только пик при задержке на τ, но также и пики, соответствующие задержкам на 2×τ, 3×τ и т.д. [6, 8]. Действительно, присутствие пика при задержке 2×τ можно рассматривать как подтверждение достоверности замера по пику при задержке τ (если такой имеется). Для проверки эффективности одного из возможных алгоритмов учёта периодичности РСФ проведён следующий эксперимент.

Использовались тестовые файлы с речью дикторов типа S [11]. В исходных файлах частота дискретизации 8 кГц. Путём интерполяции получены отсчёты с частотой следования 16 кГц. Приведённые в работе [11] данные о периодах ОТ представляют усреднённые значения, следующие через 11,25 мс. На участках с быстрыми изменениями ОТ реальные значения периодов могут существенно (больше порогового значения ξ=0,1) отличаться от приведённых усреднённых значений, поэтому использованы откорректированные данные о периодах, следующие через 4 мс. Увеличена доля вокализованных участков в исследуемом речевом материале. В разметке [11] эта доля составляет около 52%, в используемой нами разметке вокализованные участки занимают 60% общей длительности, увеличение достигнуто за счёт учёта вокализованных звуков с малой амплитудой.

При измерении периодов ОТ по положению максимума РСФ в 6,44% случаев наблюдались грубые ошибки (отличие измеренного периода от истинного более 10%). Если же выбирать два самых больших пика РСФ m1(τ1) и m2(τ2) (τ2>τ1) и при соблюдении условия |τ1-τ2/2|<δ увеличивать амплитуду пика m1(τ1) на ν×m2(τ2), после чего определять период ОТ по положению пика с максимальной амплитудой, то при таком алгоритме количество грубых ошибок измерения ОТ уменьшается до 3,69% (если τ — количество тактов частоты 16 кГц, то оптимальное значение δ=6. Оптимальное значение ν=0,31).

Исследовалась зависимость вероятности грубых ошибок от частоты среза ФНЧ, ограничивающего полосу частот, поступающих на вход измерителя. При измерении периодов ОТ по максимуму РСФ при работе с РС, содержащими низкочастотные составляющие (тестовые сигналы типа S [11]), оказалось целесообразным установить частоту среза ФНЧ равной максимальной частоте ОТ (500 Гц). При увеличении полосы анализируемых частот до 1000 Гц количество грубых ошибок увеличивается в 1,07 раза. В дальнейших экспериментах частота среза ФНЧ выбрана равной 500 Гц. Очевидно, что при работе с РС с подавленными низкочастотными составляющими частоту среза ФНЧ необходимо увеличить.

Приведённые результаты относятся к случаю измерения периодов ОТ по функции РСФ $S(t,\tau)$ для фиксированных значений t без учёта результатов измерений на соседних интервалах. Далее рассматриваются возможности повышения достоверности за счёт учёта зависимостей между периодами ОТ.

Учёт зависимостей между периодами ОТ

В качестве кандидатов на результат замера периода ОТ функции $S(t,\tau)$ для момента времени t рассматриваются восемь значений $\tau1(t),\tau2(t),\ldots,\tau8(t)$, соответствующих восеми наибольшим пикам функции $S(t,\tau)$. Каждому замеру $\tau i(t)$ присваивается вес, пропорциональный величине функции $S(t,\tau i(t))$. Простейший способ учёта результатов замеров на соседних сегментах состоит в следующем. Для каждого $\tau i(t)$ строится траектория движения ОТ по прошлым и будущим значениям и вес этих траекторий прибавляется к весу замера $\tau i(t)$, после чего выбирается замер с максимальным суммарным весом. Траектории можно строить по правилу выбора ближайших (минимум абсолютной величины разности) замеров ОТ при движении вперёд и назад. Вес траектории можно вычислять в виде суммы весов выбранных замеров, умноженной на весовой коэффициент Wk. Этот алгоритм назовём Алг. 1.

В Таблице 1 приведены результаты экспериментов. Длина траектории L определяет количество соседних замеров, учитываемых при движении в каждую из сторон. L=0 означает, что соседние замеры не используются. Коэффициенты Wk определяют веса, с которыми складываются замеры $S(t-k,\tau)$ и $S(t+k,\tau)$ при формировании суммарного веса замеров для момента t. Pe — количество грубых сбоев в процентах.

Таблица 1

Процент ошибок Pe для Алг. 1

L	0	1	2	3	5
Pe, %	3,69	2,01	1,74	1,41	1,30
Wk	—	0,60	0,38	0,32	0,12

На *рис. 6* представлен пример работы Алг.1. По оси ординат отложены значения задержки τ при измерении периодов ОТ. Значения τ задаются количеством тактов частоты 16 кГц. По оси абсцисс отложены моменты времени t, следующие через 4 мс (64 такта частоты 16 кГц). Для каждого t выбираются восемь наибольших пиков РСФ, значения τ для которых отображены цветными кружками. Цвета назначены в порядке уменьшения амплитуд пиков следующим образом: красный, жёлтый, зелёный, синий, чёрный, далее белый. Правильные значения периодов представлены красной траекторией. Выбор результата измерения ОТ производится для t=10. Результаты предыдущих выборов отображены двойной зелёной траекторией. *Рис. 6* показывает, что полученная траектория измерений ОТ совпадает с траекторией правильных значений, грубые ошибки измерения периодов ОТ, которые произошли бы при t=3 и t=8 в случае измерения путём выбора максимального пика РСФ без учёта замеров на соседних интервалах, исправляются.

На *рис. 7* представлен пример, когда обработка замеров согласно Алг. 1 не исправляет, а даже увеличивает ошибки. Однако такие случаи встречаются редко, в среднем обработка замеров по Алг. 1 позволяет уменьшить количество грубых ошибок примерно в три раза.

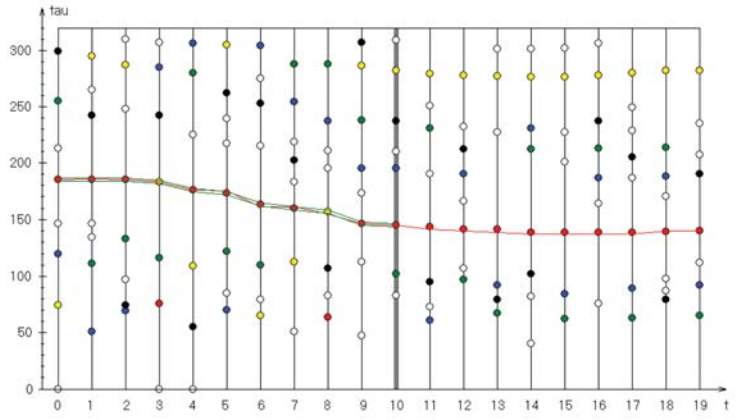


Рис. 6. Замеры ОТ и результаты их обработки. Алгоритм 1. Пример 1

Ошибки определения ОТ чаще всего случаются на начальных и конечных участках вокализованных сегментов. Здесь обычно наблюдаются значительные вариации периодов ОТ. Кроме того, при формировании траектории одна из её частей попадает на паузу или на невокализованный сегмент речи. Функция (1) инвариантна к уровню РС, поэтому на паузах иногда встречаются случайные выбросы, искажающие вес траектории. Для устранения этого недостатка при формировании веса траектории учитываются уровни сигнала на соответствующих участках траектории путём умножения значений пиков функции $S(t,\tau)$ на корень квадратный из средней амплитуды сигнала для времени t.

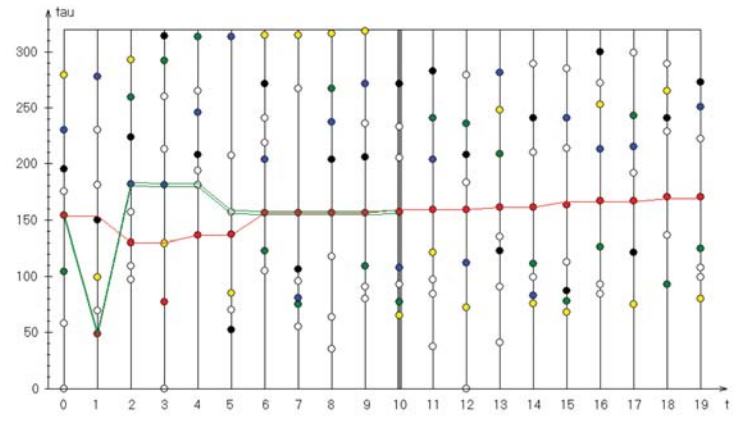


Рис. 7. Замеры ОТ и результаты их обработки. Алгоритм 1. Пример 2

В Алг. 1 при построении траекторий выбирается ближайший по τ соседний замер, но не учитывается достоверность замера (величина пика РСФ) и абсолютная величина расстояния по τ (вероятность перехода). В более совершенном Алг. 2 при построении траектории выбирается замер с максимальной вероятностью произведения достоверности измерения на вероятность перехода. В логарифмическом масштабе это сумма величины пика РСФ и логарифма вероятности перехода. Движение ОТ аппроксимируется нестационарным Марковским

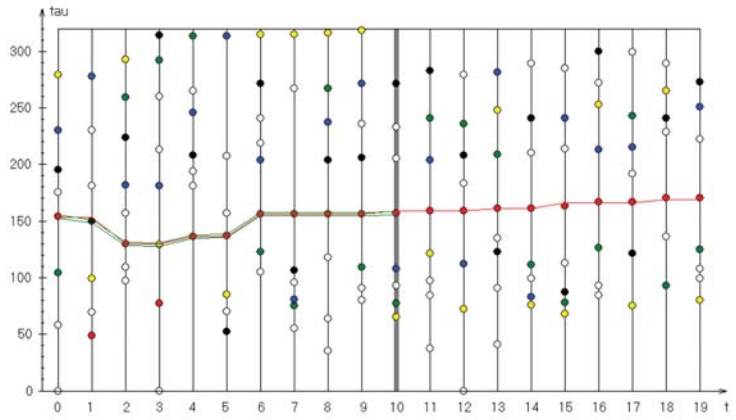


Рис. 8. Замеры ОТ и результаты их обработки. Алгоритм 2. Пример 1

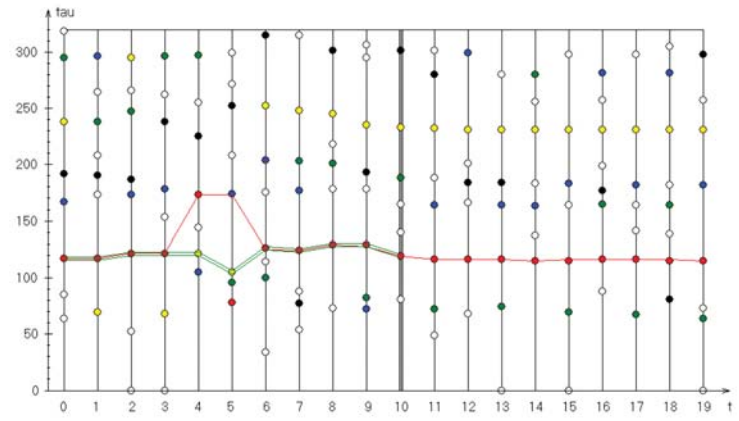


Рис. 9. Замеры ОТ и результаты их обработки. Алгоритм 2. Пример 2

процессом с функцией переходов вида $\exp(-|d\tau|/\sigma(t))$, где $d\tau$ — изменение периода ОТ на интервале между соседними замерами, $\sigma(t)$ — изменяющаяся во времени дисперсия. Дисперсия увеличивается на границах вокализованных сегментов и при резком изменении формы РС. Грубой оценкой $\sigma(t)$ могут служить вариации уровня анализируемого РС. В Алг. 2 вес траектории определяется в виде суммы амплитуд выбранных пиков РСФ и логарифмов вероятностей переходов $-|d\tau|/\sigma(t)$.

Испытания Алг. 2 при длине траектории L=9 (дополнительная задержка на 4×9=36 мс) дали 0,44% грубых ошибок. На *рис. 8* представлены результаты обработки по Алг. 2 тех же данных, что приведены на *рис. 7.* Там, где Алг. 1 даёт увеличение количества ошибок, Алг. 2 полностью исправляет ошибки. Но и Алг. 2 не исправляет ошибок во всех случаях, а иногда и увеличивает их количество. На *рис. 9* приведён пример, когда резкое изменение периода ОТ при t=4 и t=5 вызвало две грубые ошибки (при выборе максимального пика РСФ без учёта соседей была бы только одна ошибка при t=5). Тем не менее, в среднем Алг. 2 обеспечивает существенное уменьшение количества грубых ошибок. В таблице 2 приведены опубликованные в работе [14] результаты сравнения различных ВОТ с разработанным В.В. Бабкиным алгоритмом. Таблица дополнена двумя последними столбцами, где приведены данные наших исследований.

Таблица 2

Вероятности грубых ошибок Pe для различных алгоритмов выделения ОТ при отсутствии внешних шумов [14]. Тестовый сигнал типа S [11]

ВОТ	G.729.AB	G.723.1	MELP	LPC10E	HVXC	[14]	Алг. 1	Алг. 2
Pe, %	10,6	10,9	4,6	6,0	5,7	4,5	1,3	0,4

Помехоустойчивость

Работа программы Алг. 2 проверялась при наложении на тестовый сигнал S [11] помех с равномерным спектром. Испытания проводились с помехами с гауссовским распределением (ГП) и с помехами с экспоненциальным распределением амплитуд (ЭП). Отношение сигнал/шум (С/Ш) определялось как отношение средней мощности РС на вокализованных участках к средней мощности помехи. В таблице 3 приведены результаты испытаний.

Таблица 3

Вероятности грубых ошибок Ре для ВОТ Алг. 2 при воздействии помех

С/Ш, дБ	20	15	10	5
ГП, Ре, %	1,6	3,3	6,6	13,3
ЭП, Ре, %	1,0	1,6	2,7	4,8

Полученные данные уступают результатам, опубликованным в работе [14] (для помехи с гауссовским распределением Ре=6,9% при С/Ш=0 дБ). Видимо, это объясняется двумя причинами:

1. В использованной нами разметке сигнала S к вокализованным звукам отнесены сегменты РС с малыми амплитудами, которые в разметке [11] считаются паузой. При воздействии шумов ошибки возникают чаще всего на слабых сигналах.

2. Параметры Алг. 2 оптимизированы на незашумленных РС. При воздействии шумов некоторые параметры целесообразно изменить. Например, надо увеличить время накопления при вычислении РСФ. Последнее исключит возможность отслеживать быстрые изменения периодов ОТ, но средний процент ошибок при воздействии сильных шумов уменьшится.

При воздействии помех типа ЭП помехоустойчивость ВОТ существенно выше. Видимо, это объясняется тем, что алгоритм вычисления РСФ ориентирован на экспоненциальное распределение помех. В связи с этим возникло предположение, что при помехах с гауссовским распределением измерение периодов ОТ по АКФ даст существенно лучшие результаты, чем измерение по РСФ. Эксперименты не подтвердили это предположение. При измерении периодов ОТ путём выбора максимума АКФ или РСФ (без учёта замеров на соседних сегментах) при воздействии гауссовских помех количество грубых ошибок оказалось примерно одинаковым.

Заключение

Автокорреляционный метод выделения ОТ речи, предложенный проф. А.А. Пироговым 50 лет назад, в настоящее время широко используется при исследованиях речи и в аппаратуре связи и даёт хорошие результаты. Перспективы дальнейшего совершенствования метода связаны с более полным учётом особенностей речеобразования и с использованием аппарата теории статистических решений для синтеза алгоритмов обработки речевого сигнала, приближающихся к оптимальным. Точность измерений определяется уровнем внутренних шумов, которые учитывают отклонения используемой математической модели от реального РС. Нормализация по уровню (коэффициент gain в формуле (1)) позволяет уменьшить уровень шумов, обусловленных изменениями амплитуд РС на анализируемом интервале, где, согласно модели, сигнал представляется периодическим процессом с постоянными амплитудами гармоник. Следующим шагом может быть нормализация по форме РС. Резкие изменения формы речевой волны (рис. 4), обусловленные перестройками артикуляционного аппарата, затрудняют измерения периодов ОТ. Нормализация по форме, например, путём выравнивания спектра также может способствовать решению трудной проблемы выделения основного тона речи.

Литература

1. *Пирогов А.А.* Устройство для автоматического определения частоты основного тона. Реестр изобретений СССР. Авторское свидетельство №129739 с приоритетом от 08.06.1958 г. Бюллетень изобретений и товарных знаков. 1960. № 13. С. 38.

2. *Баронин С.П.* Автокорреляционный метод выделения основного тона речи: Сб. трудов Гос. НИИ Мин. связи СССР. Вып. 3(24). М., 1961. С. 93–102.

3. Coding of Speech at 8 kbit/s. ITU-T Recommendation G.729. 1996.

4. Dual Rate Speech Coder. ITU-T Recommendation G.723.1. 1996.

5. Speech Service Option Standard for Wideband Spread Spectrum Systems. ANSI/TIA/EIA-96-C. 1998.

6. *Баронин С.П.* Статистические методы анализа речевых сигналов // Электросвязь, 1966. № 5. С. 50–56.

7. *Баронин С.П.* О построении многоканальных схем выделения основного тона речи: Сб. тр. Гос. НИИ мс. 1965. Вып. 3(39). С. 17–24.

8. *Баронин С.П.* Статистические методы анализа речевых сигналов: Канд. дисс. М.: НИИР, 1968.

9. *Баронин С.П., Куштуев А.И.* О построении схем адаптации анализаторов частоты основного тона речи: Тез. докл. 7-й Всесоюзной акустической конференции. Л., 1971.

10. Вокодерная телефония / Под ред. А.А. Пирогова. М.: Связь, 1974.

11. *Бабкин В.В.* Тестовые файлы для оценки помехоустойчивости выделителей ОТ. Файл S. СПб.: Центр ЦОС СПб ГУТ, 2005.

12. *Соболев В.Н., Баронин С.П.* Исследование сдвигового метода выделения основного тона речи // Электросвязь. 1968. № 12. С. 30–36.

13. *Соболев В.Н.* Информационные технологии в синтетической телефонии. М.: ИРИАС, 2007.

14. *Бабкин В.В.* Помехоустойчивый выделитель основного тона речи. 7-я Международная конференция и выставка «Цифровая обработка сигналов и её применение (DSPA-2005)». М., 2005.

Баронин Сергей Павлович —

кандидат технических наук, ведущий научный сотрудник Государственного научно-производственного центра «ВИГСТАР». Занимается исследованиями и разработкой современных систем связи, в том числе систем эффективного кодирования и автоматического распознавания речи. Опубликовал более 100 научных работ, получил авторские свидетельства на 30 изобретений. Работы, посвящённые автокорреляционному методу выделения основного тона речи и статистическим методам анализа речи, в том числе анализу авторегрессионной модели речи (метод линейного предсказания), очистки речи от шумов путём обработки спектров, векторному квантованию дельта-кепстров (фонетической функции Пирогова) с целью сжатия информации, опубликованы ещё в 60-е годы прошлого века.

Речевые корпуса на новом технологическом витке

О.Ф. Кривнова,
доктор филологических наук

Корпуса звучащей речи, которые называют также речевыми базами данных, представляют собой важнейший тип языковых ресурсов. Интерес к созданию речевых корпусов был в значительной степени инициирован разработками в области автоматического распознавания речи, где исследователям приходится сталкиваться с огромной акустической вариативностью звуковых единиц языка, которая имеет весьма разнообразные источники. Однако сегодня речевые корпуса имеют более широкое применение, и их разработка сама по себе постепенно превращается в самостоятельное и популярное направление речевых технологий. В статье рассматривается история разработок в этой области, их современное состояние, даётся краткое описание речевых и лингвистических корпусов для русского языка.

Речевой корпус как разновидность языковых ресурсов

Корпуса звучащей речи, которые называют также речевыми базами данных, представляют собой важнейший тип языковых ресурсов.

Речевой корпус — это структурированное множество речевых фрагментов, которое обеспечено программными средствами доступа к отдельным элементам корпуса. **Речевой фрагмент** как базовая единица корпуса представляет собой оцифрованный фрагмент речевого сигнала, который сопровождается ассоциированной информацией определённого типа (типов). Такая информация называется также **аннотацией** к речевому фрагменту.

В настоящее время задача создания больших, разнообразных и информационно богатых (многоуровневых) речевых корпусов, а также удобного и надёжного инструментария для их разработки и использования становится всё более актуальной как для компьютерных приложений, так и для фундаментальных научных исследований. Современные системы распознавания речи, дающие наиболее высокие показатели надёжности, базируются преимущественно на методах статистического моделирования речевых и языковых явлений и требуют обучения на больших массивах аннотированной звучащей речи, записанной от многих дикторов (не менее 100 человек).

Современный подход к синтезу речи по тексту, основанный на конкатенации акустических фрагментов разной размерности, также предполагает использование больших речевых корпусов [1]. Попытки применения статистических методов для формирования речевого сигнала при синтезе речи также способствуют возрастанию роли речевых корпусов в дальнейшем развитии речевых технологий разного профиля [2].

Специалисты считают, что корпусной подход (corpus-based approach) является определяющим для развития технологий синтеза, особенно при моделировании просодических характеристик речи и индивидуальных особенностей говорящего. Отмечаются также такие достоинства этого подхода, как формализация процедур обучения, применение итеративного обучающего процесса с исправлением возникающих и контролируемых ошибок, возможность контроля и объективной оценки работы различных прикладных систем на стандартизованной основе (на одних и тех же речевых корпусах). Практика показывает, что при наличии речевых корпусов и технологий обучения создание прототипической версии автоматического распознавателя или синтезатора речи занимает не так уж много времени. В литературе указываются сроки от двух месяцев до полугода. Для коммерчески ориентированных разработок это немаловажное обстоятельство.

Современные речевые технологии базируются не только на речевых корпусах, но нуждаются также и в более широких, богатых информационно, лингвистических корпусах, т.е. коллекциях специальным образом обработанных текстов, как письменных, так и устных, на данном языке. Какие принципы лежат в основе устройства текстовых лингвистических корпусов, для каких речевых задач их можно использовать и как — это отдельная тема, которая в данной статье не обсуждается (см., однако, в конце статьи краткое описание национального корпуса русского языка — НКРЯ).

Было бы неправильно думать, что речевые корпуса представляют интерес только для речевых технологий. Наличие представительных речевых корпусов в электронном формате, снабжённых специальной информацией, уровень развития современных программных средств обработки звучащей речи и постоянно возрастающие мощности компьютерной техники дают учёным-лингвистам недоступную ранее возможность для проведения крупномасштабных и статистически достоверных исследований на разнообразном речевом материале. Среди других социально важных применений речевых корпусов вне сферы собственно речевых технологий можно отметить задачи обучения иностранным языкам, лингвокриминалистику и медицинскую диагностику.

Из истории разработок

Первые речевые корпусы (далее РК) были созданы в первой половине 80-х годов прошлого века в США для американского варианта английского языка, где их разработка финансировалась Министерством обороны, а организация работ была поручена национальному институту стандартов и технологий NIST (National Institute of Standards and Technology). Основное назначение первых РК — тестирование и оценка работы систем распознавания речи на одном и том же стандартном речевом материале.

Во второй половине 80-х годов произошли значительные сдвиги в компьютерной технике: возросла мощность компьютеров и объёмы хранения данных; происходило массовое внедрение персональных компьютеров. К этому времени были подведены окончательные итоги крупномасштабных государственных проектов ARPA/DARPA (Defense Department's Advanced Research Projects Agency) США, которые были направлены на анализ и оценку перспектив распознавания слитной речи с большим словарём и человеко-машинных диалоговых систем с устным вводом информации [3]. Проведённые в рамках этих проектов исследования ярко продемонстрировали преимущества систем распознавания речи на основе теории распознавания образов, статистических методов и обучающих речевых корпусов (сравнительно с экспертными системами на основе лингвистических знаний и правил). Этот временнòй период можно считать началом формирования нового направления речевых технологий, связанного с созданием речевых корпусов.

При государственной поддержке в США в 80-е годы были созданы: TI-DIGITS корпус (1984) для тестирования систем распознавания изолированных цифр и цифровых последовательностей; Road Rally для анализа и распознавания ключевых слов (word spotting); King Corpus для систем идентификации говорящего (speaker recognition); корпус TIMIT (Texas Instruments & Massachusetts Institute of Technology, Acoustic-Phonetic Continuous Speech Corpus 1980–1990), который послужил прототипом для многих других речевых корпусов, в том числе и не англоязычных. Были разработаны также специализированные речевые корпуса Resourse Management (RM) и Wall Street Journal (WSJ, позднее CSRNAB (Continuous Speech Recognition of North American Business News)) для исследований в области распознавания слитной речи и корпус Air Travel Information Service (ATIS) для исследования спонтанной речи и понимания естественного языка в диалоговых системах. Краткая характеристика перечисленных корпусов даётся ниже в таблице 1.

Таблица 1

Краткая характеристика речевых корпусов 80-х годов XX в.

Название	Назначение: для использования в проектах	Язык	Год	Общая характеристика
TI-DIGITS	Распознавание цифр и их последовательностей	Амер — англ.	1984	?
ROAD RALLY	Распознавание ключевых слов в речевом массиве	Амер — англ.	Первая половина 80-х г.	?
KING CORPUS	Идентификация говорящего	Амер — англ.		?
RM	Распознавание запросов в области военно-морской службы	Амер — англ.	Вторая половина 80-х г.	160 дикторов, словарь 1000 слов, 21000 предложений
WSJ	Дикторонезависимое распознавание слитной речи	Амер — англ.		Чтение новостных текстов в различных сферах бизнеса, словарь 20000 слов
ATIS	Распознавание запросов в сфере обслуживания гражданской авиации	Амер — англ.		Спонтанные диалоги сфере обслуживания гражданской авиации
TIMIT	Для широкого использования; дикторонезависимое распознавание слитной речи; научные исследования	Амер — англ.	80–90-е годы	630 дикторов, отдельные предложения — 2432; словарь 6229 единиц, адаптированный Merriam-Webster Pocket Dictionary 964

Практика показала, что создание хорошего речевого корпуса представляет собой довольно сложную технологическую задачу, требующую значительных финансовых и кадровых вложений. Для её решения в 90-е годы XX в. были созданы специальные координационные центры по сбору, хранению, распространению и созданию общедоступных и стандартизованных языковых ресурсов, в том числе речевых. Среди них:

- **LDC** — Linguistic Data Consorcium, http://www.ldc.upenn.edu.
- **CSLU** — Center for Spoken Language Understanding, Oregon Graduate Institute, http://www. CSLU.ogi.edu.
- **ELRA** — European Language Resources Association, http://www.icp.grenet.fr/elra.

Более подробные сведения о центрах языковых и речевых ресурсов можно найти в [5]. С момента образования указанных координационных центров начался второй этап технологического развития РК.

Речевые корпуса на современном этапе технологического развития (конец XX — начало XXI века)

Коллекции речевых корпусов, которые предлагаются координационными центрами, с каждым годом увеличиваются, и всё больше специалистов участвует в их разработке. Одновременно растёт мощность, разнообразие и программное оснащение самих корпусов.

Самым мощным на сегодняшний день является центр LDC (США), который в 2008 году отмечает свой 15-летний юбилей. За прошедшие годы центр участвовал в создании и распространении более 50 000 лингвистических, в том числе речевых, корпусов на разном языковом материале[1]. В коллекции центра около 50 речевых корпусов, содержащих сотни часов звучащей речи, а также современный компьютерный инструментарий для обработки звучащей речи и создания речевых баз данных. На сайте центра размещена карта мира, на которой обозначены региональные исследовательские центры, участвующие в создании лингвистических корпусов разного профиля. Их число постоянно растёт, и это свидетельствует об образовании особого профессионального сообщества — Linguistic & Speech database community (*рис. 1*).

Рис. 1. Карта языковых ресурсов и разработчиков LDC

Можно назвать и другие признаки перехода речевых корпусов из обязательной составляющей других речевых технологий в самостоятельное технологическое направление. Так, на рубеже веков в фокусе внимания разработчиков и других заинтересованных специалистов оказались вопросы **стандартизации** методов, представления данных, аннотаций и инструментария корпусных ресурсов. Начало широкому обсуждению этих проблем было положено выходом книги «The Handbook of Standards and Resources for Spoken Language Systems». Ed. Gibbon D., Moore R., Winski R., 1997 [4]. В области речевых корпусов долгое время образцом для разработчиков служил американский корпус TIMIT, (табл. 1), а также подробнее [5].

[1] Университетские исследователи получают значительные скидки при приобретении корпуса из коллекции LDC.

С начала XXI века по инициативе LDC проводятся регулярные рабочие совещания и конференции по разным вопросам создания лингвистических баз данных. Очень важным событием оказалась дискуссия о типах и средствах лингвистических аннотаций в корпусах разного профиля и целевого использования, которая была организована отделом аннотаций LDC при поддержке IRCS (Institute for Research in Cognitive Science), США в декабре 2001; материалы этого совещания до сих пор доступны на сайте LDC: [IRCS Workshop on Linguistic Databases [Dec 2001]. Актуальная проблематика речевых корпусов рассматривалась предварительно на страницах двух специальных выпусков журнала «Speech Communication» в начале 2001 г. [6]. Представленные здесь публикации заслуживают отдельного обсуждения, назовём лишь их тематическую рубрикацию:

- представление речевых данных, структура и содержание аннотаций;
- связи между аннотациями и сигналами;
- структура и организация баз данных;
- проблемы компьютерной разработки и использования РК;
- фундаментальные проблемы методологии исследований и разработок, относящихся к аннотированным РК.

Многие участники вышеупомянутой дискуссии затронули и целый ряд важных организационных вопросов, см. например, доклад [7] с показательным названием «Writing a Corpus Cookbook». Была отмечена, в частности, необходимость подготовки методического руководства и рекомендаций для оптимизации проектов по созданию лингвистических корпусов. Подчёркивалось также, что в рамках существующего и возрастающего разнообразия ресурсов, в том числе электронных, трудно получить информацию о том, какие ресурсы уже существуют и доступны, хотя это необходимо по практическим и этическим соображениям.

В качестве реакции на эти актуальные проблемы в 2002 г. был инициирован проект **OLAC** (the Open Language Archives Community, http://www.language-archives.org/, который имеет сервисную службу на сайте LDC. Цель проекта и портала — устранить разрыв между потенциальными пользователями, разработчиками и массой несвязанной информации о цифровых лингвистических ресурсах, накопленных к настоящему времени мировым сообществом.

OLAC — это пример международной кооперации, сообщества организаций и отдельных лиц, которые участвуют в создании всемирной виртуальной библиотеки языковых ресурсов путём:
- формирования согласованного мнения относительно наиболее успешных разработок и проектов по созданию цифровых архивов и языковых ресурсов;
- создания сети интерактивных лингвистических архивов и средств для их размещения в Интернете, поиска и доступа к ним.

Классификация речевых корпусов

Аннотированные речевые корпуса — важнейший компонент исследований в области звучащей речи. Сегодня они созданы и создаются для большого количества языков, научных дисциплин и технологий. Опыт, накопленный в области их разработки и использования, позволяет выделить ряд признаков, которые могут быть положены в основу классификации речевых баз данных и учитываться при проектировании нового РК. Укажем наиболее важные характеристики (см. также [8]):

- ***целевое использование корпуса:*** специализированные, технологические, общие (репрезентативные), учебно-иллюстративные;

- ***тип речевого материала:*** дискретная речь, непрерывная речь-чтение, спонтанная речь, специальные и естественные диалоги;

- ***тип текстового материала:*** списки слов/слогов, наборы отдельных предложений, связные тексты; монотематические или политематические;

- ***тип речевого сигнала***: лабораторная речь, офисная речь, публичная речь, телефонная речь (обычная или через мобильный телефон); радио-, теле-речь, речь в условиях естественной внешней среды, иноязычная (акцентная) речь и т.д.;

- ***тип информации, ассоциированной с речевым сигналом (аннотации):*** орфографическая запись, фонемная / фонетическая транскрипция, просодическая транскрипция, акустико-фонетическая разметка сигнала: «событийная», сегментная, просодическая, включение других типов лингвистических аннотаций и комментариев, например, об индивидуальных особенностях произношения говорящего или эмоциональной окраске речевых фрагментов;

- ***тип статистической балансировки*** звуковых единиц языка: равномерная, репрезентативная, по специальной статистической схеме;

- ***наличие и типы дополнительной сигнальной информации,*** включённой в корпус наряду с речевым сигналом: простые, мультимодальные и специальные корпуса.

Горячие точки в технологии речевых корпусов

Судя по тематике и результатам текущих конференций и рабочих совещаний, горячими точками в технологическом прогрессе РК до сих пор являются финансовое обеспечение, необходимость кооперативных усилий, обеспечение общедоступности и многопрофильности речевых корпусов, стандартизация аннотаций и информационной структуры РК; разработка компьютерного инструментария для накопления, обработки, верификации речевых баз данных с активным привлечением возможностей сети Интернет. Кроме этих «канонических» проблем, специалисты обращают внимание на необходимость создания больших, разнообразных, информационно «богатых» (многоуровневых) речевых корпусов. Отмечается, в частности, растущая потребность в просодически размеченных и аннотированных РК, корпусах эмотивной и социально дифференцированной речи. Постоянно растёт интерес к мультимодальной коммуникации и мультимодальным корпусам и базам данных с компонентом звучащей речи [9], в этом году готовится специальный выпуск журнала LRE (Language Resources and Evaluation), посвящённый проблемам моделирования мультимодальной межличностной коммуникации.

Широкие, «коммуникативные» корпуса имеют не только технологическое, но и образовательное, общекультурное значение. В связи с этим нельзя не отметить, что ещё в 90-е годы XX в. корпорация Kay Elemetrics (США-Канада, www.kayelemetrics.com), мировой лидер в производстве аналогового и компьютерного инструментария для речевых исследований, начала разработку фонетических баз данных учебно-иллюстративного профиля с аудио- и видеокомпонентами, создав коллекцию таких баз для 50 разных языков. Приведём в качестве иллюстрации фрагмент подобной базы для одного из африканских языков (*рис. 2*).

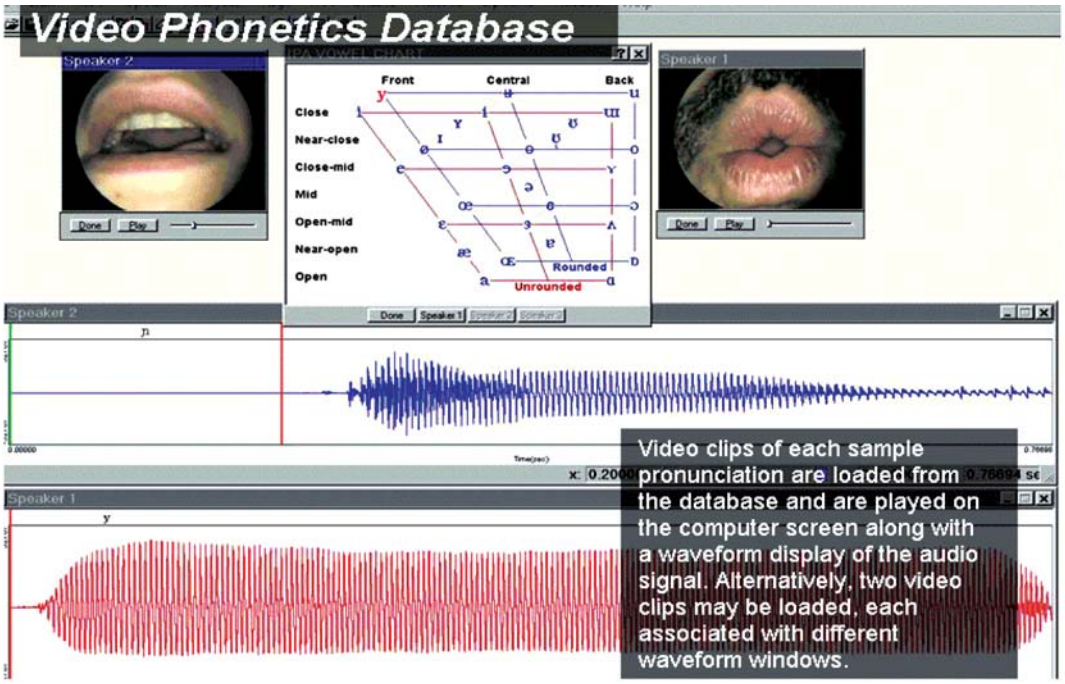


Рис. 2

Речевые корпуса для русского языка[2]

Как правило, речевые базы данных моноязычны. Речевые корпуса созданы не только для всех технологически важных языков (американского английского, немецкого, японского, китайского и др.), но и для большинства официальных языков Европейского Союза: для британского и шотландского вариантов английского языка, голландского, датского, шведского, немецкого, французского, итальянского, испанского; есть также несколько многоязычных корпусов. В результате осуществления программы Copernicus ELRA распространяет речевые корпуса и для языков Восточной Европы (польский, болгарский, эстонский, румынский и венгерский). На сайте Европейской Ассоциации в Интернете можно найти предложения и речевых корпусов для русского языка. Насколько нам известно, в их разработке принимала участие санкт-петербургская компания «Одитек» [10].

Речевой корпус ISABASE

В конце 90-х годов в Институте системного анализа (ИСА) РАН при участии специалистов речевой группы филологического ф-та МГУ был создан первый представительный речевой корпус для русского языка с разметкой речевых фрагментов на звуковые единицы. Корпус использовался не только в исследовательских целях, но и для построения автоматической системы распознавания дискретной речи [11]. Корпус моносигнальный, остальные характеристики см. ниже в таблице 2.

[2] В настоящей статье рассматриваются только те технологические речевые корпуса русского языка, которые описаны в специальных публикациях.

Таблица 2

Характеристики русского речевого корпуса ISABASE

Тип речевого материала		Дискретная речь	Дикторы/речевые фрагменты-предложения	Общий объём
Текстовый материал	**1**	Фонетически сбалансированный набор из 500 коротких предложений, монотематический	5 дикторов-мужчин и 4 диктора-женщины; 1863 фрагмента	4653 речевых фрагмента; 3713 слов
	2	Фонетически репрезентативный набор предложений, взятых из литературных текстов; политематический	15 дикторов-мужчин и 14 дикторов-женщин; 3280 фрагментов	
Типы аннотаций		Текст речевого фрагмента, фонетическ. транскрипция, ручная разметка сигнала на слова и фонемы	Транскрипционная система из 110 монофонов	

Речевой корпус RuSpeech

В 2000–2001 гг. в ИСА РАН по заказу корпорации Intel был создан также самый представительный на сегодняшний день речевой корпус русского языка **RuSpeech**, который может быть использован для разработки систем распознавания слитной русской речи [12, 13]. Общие характеристики корпуса приведены ниже в таблице 3.

Помимо самой речевой базы, важным результатом проекта ***Ruspeech*** стали отлаженная технология создания речевых корпусов и комплекс программных средств для обеспечения этой технологии. Среди последних можно отметить отладку автоматического транскриптора русской речи; создание инструментария для подготовки текстового материала с нужными фонетическими и статистическими характеристиками; создание автоматизированного рабочего места эксперта-фонетиста; программы пакетной записи дикторов; несколько программ для верификации результатов основных этапов разработки [12–15].

Укажем основные этапы проекта, которые полезно, на наш взгляд, иметь в виду потенциальным разработчикам РК:

- проектирование корпуса;
- подготовка текстового материала с возможной автоматизацией;
- подготовка фонетического обеспечения РК;
- разработка программного обеспечения для формирования речевого корпуса;
- подбор дикторского состава;
- организация записи и файлирования речевого материала;
- проверка качества записи;
- создание рабочего места эксперта-фонетиста и детальных инструкций по разметке и фонетической аннотации речевых сигналов;
- верификация аннотаций речевого материала, полученных автоматически;
- обработка результатов верификации;
- окончательное формирование и структурирование РК.

Резервы для расширения коллекции русских РК

Хотя число русских РК и баз данных с годами растёт, всё-таки это происходит очень медленно. Важным и чрезвычайно полезным резервом для будущих РК являются **фонотеки** русской звучащей речи, которые есть во многих центрах: научных (например, ИРЯ РАН им. В.В. Виноградова, ИРЯ им. А.С.Пушкина), образовательных (вузы) и культурных (музеи). Большая коллекция фонодокументов находится в Российском государственном архиве фонодокументов (РГАФД): более 200 000 единиц хранения, около 3,5 млн записей 1898–2001 гг., в том числе — восковые валики (фоновалики), оригиналы и копии 591 единиц; грампластинки, оригиналы и копии 135 000 единиц; матрицы, страховой фонд на граморигиналах 1420 единиц; тонфильмы, оригиналы и магнитные копии 1136 единиц; магнитные ленты, компакт-кассеты более 36 000 единиц; лазерные компакт-диски — более 900 единиц хранения.

Таблица 3

Дикторское и текстовое наполнение корпуса *Ruspeech*. Общие характеристики РК: моносигнальный, слитная речь, чтение

Общая характеристика	Тип речевого материала	Состав фрагментов	Дикторы/фрагменты
	Непрерывная речь; моносигнальный	50 часов записи; 30 CD, более 15 Gb; более 50 000 фрагментов-предложений	237 дикторов: 127 мужчин и 110 женщин разного возраста
Текстовый материал	1. Фонетически сбалансированный набор; политематический — нет	70 предложений, обеспечивающих полное (≥3 раз) монофонное покрытие	203 диктора: 111-м и 92-ж; каждое предложение произнесено всеми дикторами
	2. Фонетически репрезентативный (на аллофоном уровне) набор предложений, взятых из газетных и новостных текстов на интернет-сайтах; политематический	3060 предложений, обеспечивающих полное покрытие аллофонов из репрезентативного набора	203 диктора: 111-м и 92-ж по 180 предложений выборочно; каждое предложение произнесено 14 дикторами
		2000 фонетически разнообразных предложений	20 дикторов: 10-м и 10-ж по 200 предложений выборочно; каждое предложение произнесено одним диктором
Аннотации	Текст речевого фрагмента, каноническая и фактическая транскрипция, выверенная экспертами; данные о дикторе и эксперте-фонетисте — нет	Транскрипционная система из 114 монофонов	

Многие организации, обладающие фонотеками, в порядке самостоятельной инициативы проводят в настоящее время оцифровку имеющихся у них речевых материалов. Однако вопросы доступа и возможного использования этих материалов для широкого круга исследователей, и в том числе разработчиков РК, остаются открытыми и, к сожалению, даже не обсуждаются. Год русского языка, прошедший в 2007 г., никак на эту ситуацию не повлиял.

Вне пользовательской информационной зоны остаются и те РК, которые создаются по проектам, финансируемым государственными фондами РФФИ и РГНФ. Как правило, речевые базы, разрабатываемые в рамках этих проектов, где-то «исчезают» и известны узкому кругу учёных

и разработчиков по отдельным публикациям. Специалистам очевидна необходимость централизации и государственной поддержки в сфере создания, хранения, дистрибуции и обеспечения доступа к цифровым материалам русской звучащей речи.

Лучше обстоит дело с лингвистическими корпусами русского языка, важность которых трудно переоценить как для судьбы русского языка и культуры в целом, так и для научно-исследовательских работ разного профиля, хотя и здесь отмечается значительное отставание от США, Европы и Японии. С 2000 г. в России[3] ведутся работы по крупномасштабному проекту «***НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА» (НКРЯ).*** Разработка корпуса осуществляется большой группой лингвистов из Москвы, Санкт-Петербурга и других городов России в рамках программы «Филология и информатика» РАН (с частичной поддержкой Российского гуманитарного научного фонда). В Интернете в свободном доступе открыт сайт *Национальный корпус русского языка,* объёмом более 140 млн слов. Поддержка сайта и поиск по корпусу осуществляются компанией «Яндекс», здесь же на сайте можно получить подробную информацию о задачах корпуса и его текущем состоянии, см. также [16]. Мы напомним лишь, что Корпус русского языка — это собрание грамматически размеченных русских текстов XIX–XXI вв. в электронной форме, удобной для автоматического поиска и научных исследований. Практически Корпус — это информационно-справочная система, основанная на собрании русских текстов в электронной форме.

В состав Корпуса входят тексты самых разных жанров, причём в сбалансированном объёме, — произведения художественной литературы, научные, научно-популярные, религиозные и иные сочинения, публицистика, производственно-технические, юридические и многие другие тексты. НКРЯ является максимально представительным отражением русского литературного языка во всём многообразии его письменных форм. Каждому слову и каждому тексту в Корпусе приписана лингвистическая аннотация (метатекстовая, грамматическая и семантическая) на основе специального стандарта, разработанного при участии ведущих российских лингвистов.

В 2006 г. в составе Корпуса появилось несколько новых составляющих: корпус **поэтических** текстов, снабжённых, помимо обычных аннотаций, морфологических и семантических, разметкой параметров стиха — рифмы, строфики, метрики, корпус **диалектных** текстов с разметкой специфических диалектных форм, а также особый подкорпус текстов **живой русской речи,** текстов **мультимедиа** (кинофильмов) и текстов **электронной коммуникации**.

Корпус предназначен для широкого круга пользователей: профессиональных лингвистов, преподавателей русского языка, журналистов, редакторов и издателей, школьников и студентов, иностранцев, изучающих русский язык. В то же время грамматически и семантически аннотированный корпус — это не только мощное средство для многоаспектного изучения русского языка, но и важный инструмент для создания и совершенствования компьютерных средств обработки русских текстов. В частности, для речевых технологий НКРЯ создаёт возможность составления различного рода словарей и статистических моделей языка разного уровня. Правда, к сожалению, сами по себе тексты, образующие базу информационной системы НКРЯ, доступны пока что только разработчикам Корпуса. То же самое относится и к записям звучащей русской речи, на основе которых создаётся подкорпус устных текстов. Для использования этих ценных материалов в сфере русских речевых технологий необходимы специальные соглашения относительно авторских прав. В заключение остаётся выразить надежду, что такие соглашения могут быть достигнуты.

[3] О русских языковых корпусах вообще см. [17], а также электронные публикации на сайте НКРЯ www.ruscorpora.ru.

Литература

1. *Hunt A., Black A.W.* Unit selection in a concatenative speech synthesis system using a large speech database // ICASSP-96. 1996, v. 1, pp. 373–376.

2. *Black, A., Zen, H., and Tokuda, K.* Statistical Parametric Synthesis, ICASSP 2007, Hawai.

3. *Клэтт Д.Х.* Основные результаты работ по проекту ARPA // Методы автоматического распознавания речи. Т. 2. М.: Мир, 1983. С. 333–360.

4. *Gibbon, D., Moore, R., Winski, R.* (Editors) Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, 1997.

5. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С.* Речевые корпусы (опыт разработки и использование) // Труды семинара Диалог'2001 по компьютерной лингвистике и её приложениям. М., 2001.

6. Speech annotation and corpus tools // "Speech communication". Ed. S.Bird & J.Harrington, 2001, v.33, issue 1–2. http://www.ldc.upenn.edu/annotation/specom.html.

7. *Martin Wynne*. Writing a Corpus Cookbook, 2001, IRCS Workshop on Linguistic Databases [Dec 2001]. http://www.ldc.upenn.edu/annotation/ databases.html.

8. *Кривнова О.Ф.* Области применения речевых корпусов и опыт их разработки // Труды XVIII сессии Российского акустического общества РАО. Таганрог, 2006. С. 81–84.

9. LREC Workshops 2000–2008 on «Multimodal corpora: From Models of Natural Interaction to Systems».

10. *Викторов А.Б., Викторова К.О., Воронцова А.В.* и др. Речевые базы данных для задач автоматического распознавания речи и верификации говорящего // Современные речевые технологии. Сб. трудов IX сессии Российского акустического общества. М., 1999.

11. *Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я., Фарсобина В.В.* База речевых фрагментов русского языка ISABASE // Сб. «Интеллектуальные технологии ввода и обработки информации». М., Эдиториал УРСС, 1998.

12. *Богданов Д.С., Брухтий А.В., Кривнова О.Ф., Подрабинович А.Я., Строкин Г.С.* Технология формирования речевых баз данных // Сб. «Организационное управление и искусственный интеллект». М.: Эдиториал УРСС, 2003.

13. *Arlazarov V.L., Bogdanov D.S. Krivnova O. F., Podrabinovitch A. Ya.* Creation of Russian Speech Databases: Design, Processing, Development Tools // International Conference SPECOM'2004. Proceedings. S-Pb. Russia, 2004. Pp: 650–656.

14. *Кривнова О.Ф.* Фонетическое обеспечение для построения речевого корпуса // Труды XIII сессии Российского акустического общества РАО. М., 2003.

15. *Захаров Л.М., Кривнова О.Ф., Строкин Г.С.* Подбор текстового материала и статистический инструментарий для создания речевых корпусов // Труды XI сессии РАО. М., 2001.

16. Национальный корпус русского языка: 2000-2005. Результаты и перспективы. М.: Индрик, 2005; www.ruscorpora.ru

17. *Резникова Т.И., Копотев М.В.* Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов) // Национальный корпус русского языка: 2000–2005. Результаты и перспективы. М.: Индрик, 2005. С. 31–61.

Кривнова Ольга Фёдоровна —

окончила филологический факультет Московского государственного университета им. М.В. Ломоносова по специальности «структурная и прикладная лингвистика». Работает на филологическом факультете МГУ им. М.В. в должности старшего научного сотрудника. Доктор филологических наук, имеет звания «Старший научный сотрудник», «Заслуженный научный сотрудник Московского университета». Член Фонетической комиссии при ОЛЯ РАН, секции «Акустика речи» Российского акустического общества, редколлегии периодического издания «Проблемы фонетики» ИРЯ РАН , редколлегии журнала «Речевые технологии». Имеет более 100 печатных работ.

Классификация звуков русской речи с помощью бинарных решающих деревьев

В.Б. Кузнецов,
кандидат филологических наук

В.Я. Чучупал,
кандидат физико-математических наук

Рассматривается вопрос контекстно-зависимой классификации звуков русской речи с помощью построения бинарных решающих деревьев. В качестве речевого материала использовалась обучающая выборка базы данных TeCoRus, предназначенная для приложений, использующих телефонный канал связи. В работе приводится описание результатов экспериментальной классификации русских гласных и согласных в зависимости от контекста.

В последние годы традиционная фонетика претерпевает значительные изменения под напором идей, методов, инструментария и достигнутых результатов в области речевых технологий. В частности, это относится и к такой проблеме, как спектральная классификация звуков речи. Ведётся острая полемика между теми исследователями, которые считают, что фонетическое качество гласного определяется его формантной структурой на стационарном сегменте, где достигается или по крайней мере осуществляется максимальное приближение к акустической цели данного звука, и теми учёными, для которых характер гласного задаётся динамикой его спектральных параметров.

Проблема инвентаризации звуков речи приобрела на сегодняшний день исключительное значение прежде всего в речевой технологии [4]. Так, решая задачу синтеза речи путём «склеивания» сегментов, соизмеримых с отдельным звуком, нужно определить, сколько реализаций звука [а] необходимо взять, чтобы обеспечить приемлемое качество звучания. Если не учитывать фонотактические ограничения, то только для одного ударного [а] необходимо рассмотреть более 1000 различных контекстов. При построении систем распознавании речи также важно располагать оптимальным инвентарём звуков, для которых будут строиться акустические модели. Оптимальность здесь

подразумевает, в частности, компромисс между количеством акустических моделей, точностью представления речевого материала и возможностью оценки параметров моделей на доступных обучающих выборках.

Решение этого вопроса на практике приводит к необходимости ввести понятие обобщённого (типичного) аллофона. В работах создателей системы синтеза русской речи (филологический факультет МГУ) обобщённый аллофон понимается как «акустически и перцептивно различаемая контекстная реализация фонемы» [1]. Формирование множества обобщённых аллофонов осуществляется экспертным путём на основе акустико-фонетических знаний. В одном из последних вариантов синтеза число обобщённых аллофонов гласных достигало 1100, а при разработке тем же коллективом исследователей речевой базы данных предполагается, что в ней будет содержаться как минимум 1800 аллофонов.

Очевидно, что экспертный подход к определению инвентаря обобщённых аллофонов наряду с достоинствами обладает рядом существенных ограничений. Во-первых, принятие экспертом решений во многих случаях не удаётся формализовать (особенно это относится к перцептивным оценкам), во-вторых, объём акустико-фонетических знаний — величина непостоянная, в третьих, эксперт не в состоянии провести на высоком уровне сопоставительный слуховой и акустический анализ большого числа аллофонов.

Альтернативой экспертному подходу могут служить методы вероятностного моделирования, играющие сегодня ведущую роль в автоматическом распознавании и синтезе речи. В настоящей работе для моделирования спектральной динамики гласных используется скрытая марковская модель [6], которая позволяет представить звук в виде последовательных состояний, соотносимых с членением звука на сегменты (субаллофоны). В нашем случае гласный разделяется на три отрезка одинаковой длины (начальный и конечный формантные переходы плюс вокалическое ядро). В качестве алгоритма классификации состояний СММ применяется метод кластеризации сверху-вниз бинарного решающего дерева [10]. Суть этого метода применительно к нашей задаче заключается в следующем. На первом шаге построения дерева акустические наблюдения (вектора из параметров) всех звуков объединены в одном корневом узле, для которого строится общая акустическая модель. Затем из заранее сформированного списка бинарных вопросов (то есть вопросов, которые допускают только два ответа — да и нет), которые могут относиться как к самому гласному, так и окружающим его звукам, выбирается вопрос, дающий наилучшее в некотором смысле разбиение множества наблюдений корневого узла на два дочерних. На следующем шаге среди всех возможных пар «узел-вопрос» ищется та, что обеспечивает последующее оптимальное ветвление дерева. При выполнении определённых условий построение дерева считается завершённым, а листья этого дерева (терминальные узлы) являются искомыми классами субаллофонов, из которых конструируются исходные аллофоны.

Речевой материал и модель звука

В качестве речевого материала использовалась обучающая выборка базы данных TeCoRus [9], предназначенная для приложений, использующих телефонный канал связи. Обучающая выборка представляет собой шестичасовую запись чтения шестью дикторами (из них 3 женщины) фонетически представительного множества 510 отдельных предложений, отсегментированных вручную. В экспериментах по классификации гласных использовались записи только дикторов-мужчин.

В качестве параметрического описания речевого сигнала выбрана наиболее распространённая сейчас система признаков — мел-кепстральные коэффициенты и их первые производные. Эти параметры оценивались на 25 мс окне анализа. Вычислялось 16 коэффициентов и их первых производных. Таким образом, элементарное наблюдение представляло собой вектор из 32 параметров. Последовательность таких векторов использовалась для построения двух кодовых книг, отдельно для мел-кепстральных коэффициентов и их производных. В качестве кодовой книги применялась нейронная сеть — трёхмерная карта признаков Кохонена из 1000 элементов. Фонетическая модель звука представляла собой дискретную скрытую марковскую модель, обычно из трёх состояний.

Бинарные решающие деревья

Во многих случаях дискретной классификации или распознавания образов типичной задачей является оценка значения некоторого (конечного и дискретного) параметра по имеющимся наблюдениям. Решающее дерево — это граф, который задаёт соответствие между наблюдениями и искомыми значениями параметров. Листья решающего дерева соответствуют возможным значениям параметров, а ветви — некоторым комбинациям признаков, в соответствии с которыми наблюдения группируются в различные классы. Таким образом, решающее дерево можно рассматривать как представление алгоритма классификации наблюдений. Если каждый узел решающего дерева имеет ровно два потомка, тогда дерево называется *бинарным решающим деревом*.

Построение бинарного решающего дерева для классификации звуков речи состоит из следующих этапов [5]:

- формирование набора потенциальных корневых узлов;
- создание множества вопросов к звукам, их левым и правым контекстам, идентифицирующих их принадлежность к классу звуков или к конкретному фону;
- определение критериев ветвления узлов, включая оценку приращения логарифма коэффициента правдоподобия и минимальное количество наблюдений (заселённость) в терминальном узле (классе).

Построение общего дерева для звуков (например, всех гласных или всех согласных) начинается с единственного корневого узла, в котором без учёта контекста и состояний СММ объединены все наблюдения из выборки. Для узла q строится вероятностная модель распределения параметров наблюдений $q(x)$ и оценивается её качество. В данном случае выборка наблюдений для узла разбивалась на два равномощных подмножества, на одном из которых вычислялись эмпирические частоты распределения параметров, другое подмножество (контрольная выборка) служило для оценки качества модели, которое определялось как логарифм правдоподобия контрольной выборки относительно модели:

$$L(q) = \sum_{x_1, x_2, \ldots, x_N} \log q(x) = \sum_{x \in \Omega_q} \overline{q}(x) \log q(x),$$

где $x_1, x_2, \ldots, x_N$ — список наблюдений, составляющих контрольную выборку, а $\overline{q}(x)$ — эмпирическая вероятность появления наблюдения x в контрольной выборке Ω_q.

К корневому (родительскому) узлу ищется оптимальный вопрос из конечного множества вопросов, обеспечивающий такое расщепление родительского узла на два

дочерних, которое даёт максимальное приращение оценки качества моделирования. Расщепление узла означает, что принадлежащие этому узлу векторы параметров разделяются на два подмножества в соответствии с тем, удовлетворяют они или нет поставленному вопросу.

Пусть в результате применения некоторого вопроса узел *q* стал родителем для узлов *r* и *r'* . Мера пригодности вопроса для узла *q* определялась как величина приращения качества моделирования, то есть:

$$\Delta L(q) = L(r) + L(r') - L(q) \ ,$$

где $L(q)$ — качество модели родительского узла, $L(r)$ и $L(r')$ — качество модели первого и второго дочерних узлов.

На каждом последующем шаге для текущих терминальных узлов ищется такая пара «узел-вопрос», которая обеспечивает максимальное значение .

Если найденная величина $\Delta L(q)$ превышает заранее заданное пороговое значение и число обучающих фреймов в потенциальных узлах соответствует критерию минимальной заселённости, родительский узел расщепляется на два дочерних.

Когда ни один из терминальных узлов не может быть расщеплён (например, выигрыш от расщепления данных в узле становится меньше пороговой величины или заселённость в терминальном узле становится ниже допустимого минимального значения) или число терминальных узлов достигает заранее установленного порогового значения, процедура ветвления останавливается и дерево считается построенным.

Одно из основных преимуществ, связанных с применением процедуры кластеризации «сверху-вниз» на базе решающих деревьев, состоит в том, что при классификации аллофонов, не представленных в обучающей выборке, мы можем справиться с этой ситуацией, привлекая экспертные знания о классах фонетически близких аллофонов, для которых на этапе обучения уже были получены соответствующие статистические модели.

Таким образом, задача фонетиста состоит в том, чтобы определить классы звуков, которые оказывают на своих соседей в речи сходное коартикуляционное воздействие, и выразить эти экспертные знания в форме множества бинарных вопросов (требующих ответа «да» или «нет»), которые затем будут использованы для расщепления узлов дерева.

Классификация гласных звуков

Для системы русских гласных, известных своей высокой контекстуальной вариативностью и степенью редукции, было предложено в качестве потенциальных корневых узлов 53 иерархически организованных класса. На вершине классификации находится класс «Все гласные»; классы низших уровней могут состоять как из единичных элементов (например, ударный [о]), так и из группы схожих гласных. Большое количество классов объясняется, в частности, тем, что их исходное число было практически удвоено, чтобы учесть назализацию гласных в соседстве с носовыми согласными. Результаты сегментации базы данных показали, что назализация, не являясь в русской речи смыслоразличительным признаком, регулярно проявляется в речи дикторов.

Согласные были разделены на 29 классов, в ряде случаев пересекающихся.

К самому узлу дерева (центральному элементу трифона) могло быть задано 57 вопросов. Два из этих вопросов идентифицировали принадлежность наблюдений в узле одному из состояний СММ.

Вопросы к левому и правому контексту были идентичными: 40 вопросов проверяли принадлежность звуков к широким фонетическим классам и 58 вопросов идентифицировали конкретный звук (заметим, что взрывные и аффрикаты трактовались в настоящем исследовании как сочетание двух отдельных звуков — смычки и взрыва). Последний тип вопросов был ориентирован на те случаи, когда отдельный звук был представлен в данном контексте достаточным количеством фреймов в обучающей выборке.

При построении дерева решений использовались следующие пороговые величины: минимальное приращение логарифма правдоподобия — 6.0, минимальная заселённость терминального узла ≥ 150 фреймов.

Результаты

На первом шаге построения дерева (*см. рис. 1*) основой для расщепления корневого узла «Все гласные» послужило не фонетическое качество гласного или характер контекста, а противопоставление конечной трети любого гласного его предшествующей части. В предварительном эксперименте, когда обучающая выборка была увеличена на три часа за счёт привлечения речевого материала, записанного тремя дикторами-женщинами, разбиение корневого узла произошло аналогичным образом.

Фрагмент дерева решений (верхушка)

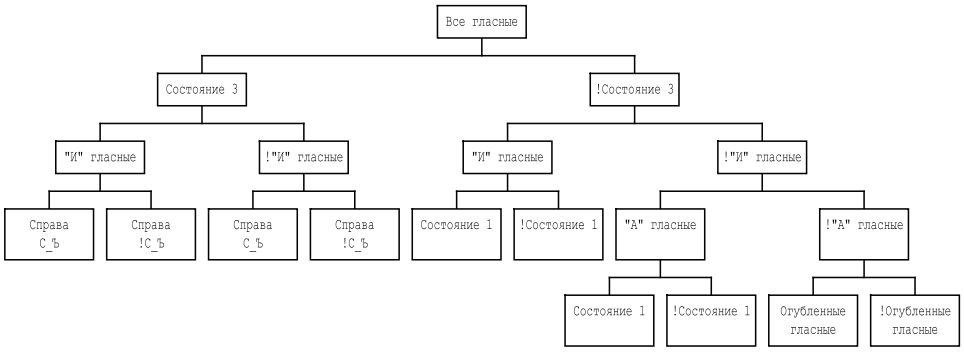


Рис. 1. Фрагмент построения глобального дерева решений (начальные шаги)

На рисунках восклицательный знак перед словом означает логическое отрицание, т.е. выражение «!Состояние 3» равноценно «Состояния 1 и 2»; С_ъ — означает любой твёрдый согласный.

На следующем шаге узлы состояний в свою очередь были расщеплены на узлы: [и]-образные гласные и все остальные. В класс [и]-образных гласных входят ударные и безударные аллофоны фонем /и/ и /е/, а также акустически нечленимые

сочетания безударных гласных, как, например, в окончании слов «Эстонии», «многие» и т.п. Последующее разделение узлов ветви «Состояние 3» зависело от твёрдости/мягкости согласного справа.

Продолжение формирования ветви, исходящей из родительского узла с характеристикой: «Состояние 3», [и]-образные гласные, правый контекст — любой твёрдый согласный», — представлено на *рис. 2*. Для краткости в описаниях узлов во всех случаях опущено, что вопрос, относящийся к согласному, адресован к правому контексту. Отметим, что один из применённых вопросов идентифицирует качество самого гласного, а именно, является ли он ударным [е]. Прямоугольники, нарисованные пунктирными линиями, являются терминальными узлами.

Построение дерева было остановлено по критерию приращения коэффициента правдоподобия. Заметим, что и зафиксированная минимальная величина заселённости классов (155 фреймов) приблизилась к критическому значению. Результирующее дерево имело 156 терминальных узлов[1]. По состояниям СММ они распределились практически равномерно: 46 узлов — «Состояние 1», 43 узла — «Состояние 2», 48 узлов — «Состояние 3» и в 19 случаях терминальный узел был построен на первых двух состояниях совместно. В последнем случае большинство элементов этого множества составляли гласный [ы] и безударные гласные после твёрдого согласного, традиционно транскрибируемые с помощью знака [ъ], а также ряд назализованных гласных.

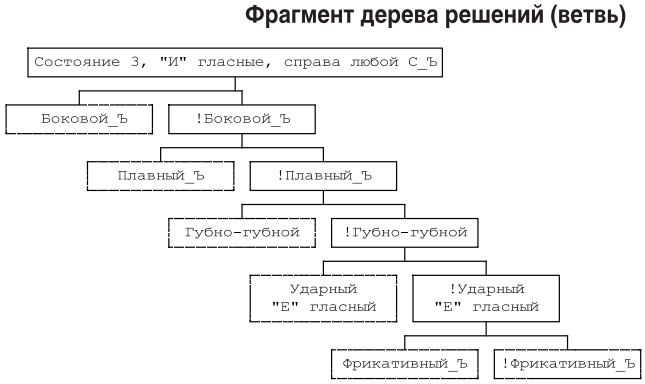


Рис. 2. Фрагмент построения глобального дерева решений (ветвь для родительского узла «Состояние 3», [и]-образные гласные, правый контекст — любой твёрдый согласный»)

В ходе построения дерева было использовано 64 различных вопроса, употреблённых в сумме 155 раз. 77% заданных вопросов относились к окружению гласного.

Левый контекст оказался значимым в 64 случаях при использовании 29 вопросов, из которых 16 относились к твёрдым согласным (39 употреблений), 12 — к мягким (24 употребления) и один — к классу [а] образных гласных (одно употребление).

Правый контекст оказался определяющим в 56 случаях при использовании 20 разных вопросов, из которых 12 относились к твёрдым согласным (41 употребление), пять — к мягким (девять употреблений) и три — к классам [а, у, е]-образных гласных (три употребления). Причём в семи случаях вопросы к левому контексту проверяли наличие конкретных звуков: [в, м, д, т, н, з', р',]; в правом контексте идентифицировался звук [с'].

[1] При увеличении критического значения приращения коэффициента правдоподобия до 11.0 построение дерева было завершено, когда число терминальных узлов равнялось 102. Подробное описание полученного инвентаря субаллофонов дано в [3, 8].

Качество самого гласного оказалось значимым в 26 случаях при использовании 13 вопросов. Эти вопросы относились как к классам гласных, так и отдельным звукам. Так, например, оказалось важным, является ли гласный [а]-образным ротовым или назализованным, или же ударным [а]. Заметим, что с назализованными гласными было связано четыре вопроса. В девяти случаях идентифицировалось состояние гласного: один раз «Состояние 3» и восемь раз «Состояние 1».

Как и ожидалось, для узла «Состояние 3» релевантными были вопросы к правому контексту. Только в одном случае был использован вопрос о мягкости левого согласного для класса огубленных гласных, справа от которых находился твёрдый согласный. В результате был образован соответствующий терминальный узел. Для узла «Состояние 1» вопросы адресовывались за редким исключением к левому контексту, и для узла «Состояние 2» имели значение вопросы к обоим контекстам.

Терминальные узлы с характеристикой «Состояние 1» и «Состояние 2» распределены между классами гласных следующим образом. Наибольшее число узлов приходится на гласный [а]: 15 на «Состояние 1» и 14 на «Состояние 2». Причём преобладают [а] в ударном и первом предударном слогах.

Следующие по числу занимаемых узлов идут [и]-образные гласные (передние, средне-верхнего подъёма): 13 — на «Состояние 1» и 10 — на «Состояние 2». В последнем случае из [и]-образных гласных исключены ударные [е], которые образовали три отдельных узла на «Состояние 1» и один на «Состояние 2». На [ы]-образные гласные пришлось по два узла на «Состояние 1» и «Состояние 2» и 14 на их объединение.

Огубленные гласные образовали на «Состояние 1» 11 узлов и на «Состояние 2» — семь узлов.

Следует упомянуть ещё об одном классе гласных, на который пришлось пять узлов «Состояние 1», шесть узлов — «Состояние 2» и четыре на оба состояния совместно. Этот класс — фонетически неоднородный и образуется в основном назализованными гласными, из которых в ряде контекстов исключаются огубленные назализованные.

По сравнению с первыми двумя состояниями классификация гласных на последнем состоянии характеризуется с традиционной точки зрения большей неопределённостью. Так, например, 16 узлов «Состояние 3» приходятся на такой класс, как «все гласные, за исключением [и]-образных». Ещё один красноречивый пример, в некотором смысле противоположный первому: два узла образованы для всех гласных, за исключением [и]-образных, [а]-образных, огубленных гласных, назализованных [а]-образных, назализованных [и]-образных и назализованного ударного [о]. В остатке имеем ротовые и назализованные [ы]- и [у]-образные гласные в ударных и безударных слогах.

На огубленные гласные приходится восемь узлов «Состояние 3», столько же — на [а]-образные.

Следует особо отметить, что в 37 случаях терминальные узлы были образованы для отдельных гласных. В частности, 27 узлов принадлежали [а], находящемуся в первом предударном и/или в ударном слогах.

Наибольшее значение для расщепления узлов дерева имели следующие признаки согласных звуков: твёрдый/мягкий, боковой, плавный, носовой, губно-губной, глухой/звонкий и фрикативный.

Обсуждение и выводы

Как показывают полученные результаты, предложенный метод представления спектральной динамики гласных в виде комбинации субаллофонов оказался достаточно эффективным. Классификация субаллофонов опирается как на чисто акустические параметры, так и на экспертные фонетические знания. Обращает на себя внимание высокая пластичность полученных классов. С одной стороны, это могут быть очень широкие классы, как, например, класс: «Состояние 3», все гласные, кроме [и]-образных, с другой стороны, несколько терминальных узлов были образованы только для ударного [а]. В ряде случаев «Состояние 1» и «Состояние 2» объединялись в одном классе.

Полученные результаты опровергают традиционное представление о том, что для характеристики гласного основное значение имеет левый контекст. Во-первых, корневой узел «все гласные» был разделён вопросом о принадлежности векторов параметров «Состоянию 3», свойства которого определяются правым контекстом. Во-вторых, количество терминальных узлов для «Состояния 1» и «Состояния 3» практически совпадает: 46 и 48. Число вопросов к левому и правому контексту — величины одного порядка: 64 и 56 вопросов соответственно.

Образование нескольких классов для назализованных гласных подтверждает целесообразность использования этого признака для характеристики русских гласных.

Классификация согласных звуков

Классы звуков и набор бинарных вопросов. Согласные звуки были представлены 48 классами, которые могли пересекаться, и 59 звуками, в число которых входили смычки и взрывы, а также вокалический компонент и удар вибранта. Классы согласных формировались как с учётом, так и без учёта признака «твёрдость/мягкость». В отдельный класс вошли сегменты разметки, указывающие на границы предложения (например, пауза, вдох и т.п.). Число классов гласных звуков было равно 25.

Общее число проверяемых вопросов составило 373. Из них 107 были адресованы к центральному элементу трифона (к узлу дерева), остальные вопросы — к левому и правому контексту.

Как и в случае классификации гласных звуков при построении дерева решений использовались следующие пороговые величины: минимальное приращение логарифма правдоподобия — 6.0, минимальная заселённость терминального узла $\geq$ 150 фреймов.

Результаты

Характеристика терминальных узлов (листьев) дерева. Построенное дерево имело 192 терминальных узла. В 142 случаях в узле содержался только один элемент (аллофон), в 26 случаях —

2 элемента и в 24 случаях число аллофонов было ≥ 3. Только правый контекст учитывался 34 раза, только левый — 40 раз, оба контекста оказались одновременно значимыми в 114 случаях. Независимыми от контекста были четыре взрыва смычных согласных: [P', B, K', G]. Более 30% аллофонов пришлось на фрикативные согласные, две трети которых были твёрдыми.

Второй по численности группой (более 25%) оказались глухие и звонкие смычки, а также удары вибранта. Доли взрывов, носовых и плавных колебались в районе 10%. Как и в случае с фрикативными согласными преобладали твёрдые аллофоны. В классы, состоящие из двух элементов, как правило, попадали фонетически близкие звуки: например, мягкие плавные [L', R'], йот и один из мягких плавных [J' + L'/R'], фрикативный и шумовой компонент аффрикаты [S, TS] и т.п.

Процесс построения дерева не противоречил принципам классификации, характерным для традиционной фонетики. Сначала весь массив наблюдений (класс «все согласные») был разбит на мягкие и немягкие согласные, которые, в свою очередь, были затем разделены на звонкие и глухие, а от группы твёрдых и мягких звонких были отделены соответствующие носовые. Начальная фаза формирования дерева представлена на *рис. 3*. Восклицательный знак обозначает логическое отрицание, апостроф после знака транскрипции обозначает мягкость согласного.

Анализ применённых вопросов. В ходе построения дерева было использовано 78 различных вопросов. Всего к центральному элементу было адресовано 49 вопросов (30 разных). Вопрос о мягких согласных был задан 20 раз, о твёрдых — 25. В 18 вопросах речь шла о твёрдых и мягких губно-губных и губно-зубных согласных.

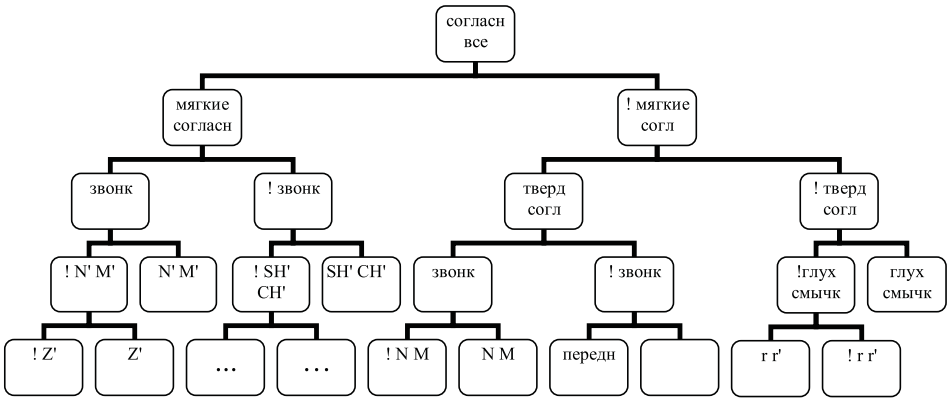


Рис. 3. Начальный этап построения дерева (обозначения см. в тексте)

Вопросы к левому контексту были использованы всего 78 раз (24 разных вопроса). В 41 случае в качестве левого контекста выступали согласные звуки и граница предложения (девять раз). Причём преобладали вопросы о твёрдых согласных. В 37 случаях левый контекст определялся гласными. Разных вопросов было пять. Класс и-образных гласных использовался 18 раз, класс а-образных гласных — 12 раз и огубленные гласные встретились пять раз.

К правому контексту распределение вопросов было следующим. Всего было применено 64 вопроса. Вопросов о согласных и границе было 53 (20 разных). В 21 случае в качестве правого контекста мог выступать любой согласный. Твёрдые согласные встретились 10 раз, мягкие — восемь раз и граница предложения — семь раз. 11 раз в качестве правого контекста выступали гласные: пять раз — огубленные гласные, четыре раза — класс а-бразных гласных.

Обсуждение и выводы

Оценивая в целом полученную классификацию согласных, в первую очередь следует отметить, что только около 10% терминальных узлов состояли из нескольких элементов. В подавляющем большинстве случаев разбиение доходило до конкретного аллофона. Последовательность построения дерева решений хорошо согласуется с преставлениями фонетистов о значимости классифицирующих признаков. Вполне ожидаемо, фрикативные согласные по числу аллофонов оказались на первом месте и аллофонов твёрдых согласных было больше, чем мягких.

Однако с точки зрения традиционной (академической) фонетики трудно объяснить тот факт, что вторую группу по численности составляют глухие и звонкие смычки, а также удары вибранта. Дело в том, что реальные условия записи речевого материала не обеспечивали получение идеальных характеристик смычек: отсутствие сигнала для глухих или наличие только «голосовой полосы» для звонких смычек. Запись дикторов проходила в тихой обстановке в обычном рабочем помещении, что делало неизбежным присутствие эффектов реверберации в записанном материале. На *рис. 4* представлена спектрограмма и спектральные срезы фрагмента слова «апатиты», стоящего в начале предложения. Спектральные срезы были сделаны с длиной окна анализа 25 мс. Спектры были получены методом БПФ и ЛПК (гладкая спектральная огибающая). Можно видеть, что спектральные характеристики паузы (соответствующее окно анализа выделено на спектрограмме белыми вертикальным линиями) существенно отличаются от спектра смычки [p] и второй смычки [t]. Причём спектр смычек в значительной степени определяется характером предшествующего гласного. Чтобы убедиться в этом, достаточно сравнить приведённые спектральные срезы для гласного [A] и последующей смычки [p]. Можно видеть, что местоположение двух первых спектральных максимумов практически идентично. Не вызывает также сомнения и-образный характер спектра смычки [t]. Целесообразность дифференциации смычек в зависимости от контекста должна быть проверена в дополнительных экспериментах по распознаванию речи.

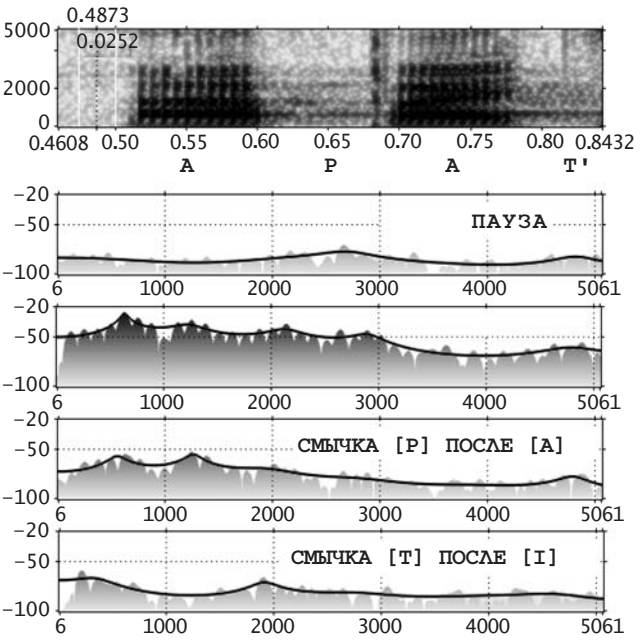


Рис. 4. Спектрограмма фрагмента слова «апатиты» и спектральные срезы, показывающие зависимость акустических характеристик глухих смычек от контекста

Как и в случае классификации гласных звуков [2, 3], для согласных оказался значимым как левый, так и правый контекст. Обращает на себя внимание тот факт, что в качестве левого контекста гласные выступили 37 раз, а в качестве правого только 11. Причём если в левом контексте преобладали и-образные гласные, то в правом они отсутствовали.

Следует отметить положительный результат, который дало отнесение взрыва аффрикат к фрикативным согласным. Результаты настоящего исследования подтверждают обоснованность использования в конкатенативном синтезе отдельных аллофонов или дифонов, когда их ближайшим соседом слева или справа оказывается пауза.

В заключение несколько общих соображений о проведённом исследовании. В настоящий момент не представляется возможным оценить устойчивость и типичность полученной классификации. Она жёстко привязана к конкретному речевому материалу, способу параметризации речевого сигнала, варианту реализации СММ, набору бинарных вопросов, критериям построения дерева решений и т.д. Проведение исчерпывающих исследований значимости этих факторов маловероятно по причине их высокой трудоёмкости. Так, в настоящей работе для построения дерева решений потребовалось около 12 часов непрерывной работы компьютера средней мощности.

Однако есть возможность постепенного сбора необходимых данных. Дело в том, что на этапе обучения распознающей системы построение классификаций аналогичных нашей, является обычным делом. Но для специалистов по речевой технологии этот результат самостоятельной ценности не представляет и, как правило, не комментируется. Ситуация может измениться, если фонетисты сами проявят инициативу, чтобы получить доступ к этим данным. Потребность во всестороннем анализе этих результатов исключительно высока. По мнению крупнейших учёных в области фундаментальной и прикладной фонетики, Г. Фанта [7] и Б. Линдблома [10], фонетическая вариативность речи огромна, но не случайна и в значительной мере систематична. Необходимо длительное и кропотливое исследование реализации звуков речи во всевозможных контекстах, чтобы по мере накопления данных удалось в конце концов, используя информацию о контексте в самом широком смысле, расклассифицировать фонетическую вариативность и снять её кажущуюся неопределённость. Тогда-то и проявится зашифрованная структура речевого сигнала.

Работа выполнялась при поддержке проектов РФФИ № 07-01-00657а и № 06-08-01534а.

Литература

1. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С.* Речевые корпусы (опыт разработки и использование) // Труды международного семинара «Диалог'2001 по компьютерной лингвистике и её приложениям». Аксаково, 2001.

2. *Кузнецов В.Б.* О принципах акустической классификации русских гласных // Язык и речь: проблемы и решения. М.: МГУ, 2004. С. 100–117.

3. *Кузнецов В.Б., Чучупал В.Я.* Инвентаризация гласных аллофонов русской речи методом кластеризации сверху-вниз бинарных деревьев решений // Акустика речи. Медицинская и биологическая акустика: Сб. т. XIII сессии Российского акустического общества. Т 3. М.: Геос, 2005. С. 54–57.

4. *Потапова Р.К.* Речь: коммуникация, информация, кибернетика. М.: Радио и связь, 1997.

5. *L.Breiman, J.Friedman* etc. «Classification and Regression Trees», Wadsworth, Inc, 1984.

6. Речевая связь с машинами (тематический выпуск) // ТИИЭР. 1985. Т. 73. № 11.

7. *Fant G.* On the speech code // Speech, Music and Hearing, KTH, Stockholm, Sweden, TMH-QPSR, Vol. 42, 2001, p. 61–72.

8. *Kouznetsov V., Chuchupal V.* Increasing trainability of ASR system by means of top-down clustering procedure based on decision trees (vowel data for Russian) // Proc. Intern. Workshop "Speech and Computer", SPECOM'04, St.-Petersburg, 2004, p. 289–291.

9. *Kouznetsov V., Chuchupal V., Makovkin K., Chichagov A.* Design and implementation of the Russian telephone speech database. // Proc. Intern. Workshop "Speech and Computer", SPECOM'99, Moscow, 1999, p. 179–181.

10. *Lindblom B.* Developmental origins of adult phonology. The interplay between phonetic emergents and the evolutionary adaptations of sound patterns // Phonetica Vol. 57, No. 2–4, 2000, p. 5–30.

11. *Nakajima Sh., Hamada H.* Automatic generation of synthesis units based on context clustering // Proc. ICASSP-88. 1988, Avr. N.Y. p. 659–662.

В.Б. Кузнецов —

профессор кафедры прикладной и экспериментальной лингвистики Московского государственного лингвистического университета. Сфера научных интересов — фундаментальные исследования процессов речепроизводства и восприятия речи (преимущественно на материале гласных звуков русского языка) и приложения в речевых технологиях (синтез речи, распознавание языка сообщения, распознавание речи). Автор более 100 научных и научно-методических публикаций, в том числе монографий «Автоматический синтез речи. Алгоритмы преобразования «буква-звук» управление длительностью речевых сегментов» (1989) и «Лингвистическое обеспечение систем синтеза речи по правилам: достижения, проблемы и перспективы» (1992).

В.Я. Чучупал —

закончил МГПИ им. В.И.Ленина в 1976 году по специальности «математика», в 1983 году закончил очную аспирантуру ВЦ АН СССР, научный руководитель — В.Н. Трунин-Донской, с тех пор работает в ВЦ РАН. Основная область интересов: распознавание и обработка речевых сигналов. Кандидат физико-математических наук, ведущий научный сотрудник.

Система автоматического распознавания языков на основе гауссовских и авторегрессионных моделей

Ю.С. Иващенко

Д.А. Леднов,
кандидат технических наук

Н.А. Любимов

Под системой автоматической идентификации языков (Automatic Language Identification — LID) подразумевается такая система, на вход которой поступают записи речевых сообщений, а на выходе формируется заключение о языке сообщения.

Необходимые требования к подобной системе — текстонезависимость и дикторонезависимость. Текстонезависимостью является устойчивость работы системы по отношению к изменению содержания входного сообщения. Под дикторонезависимостью устройства подразумевается способность системы распознавать язык сообщения, сказанного произвольным диктором.

В основе системы автоматической идентификации языков (САИЯ) лежит создание набора различных языковых моделей. Формирование моделей производится на основе фонетико-лингвистических [2] или акустических параметров речи [1]. Первый подход достаточно эффективен в распознавании, но требует много вычислительных ресурсов, что делает его неприменимым для большинства систем реального времени. В настоящей статье рассматривается второй подход, использующий акустическую параметризацию речи.

Цель работы — представить сравнительный анализ трёх различных типов акустических моделей языка.

Алгоритмы предобработки

В работе используются два различных типа предварительной обработки речевых данных. Векторы наблюдения были сформированы на основе:

- Логарифмически масштабированных кепстральных коэффициентов *(Mel Frequency Cepstral Coefficients — MFCC).*
- Смещённого дельта кепстра *(Shifted Delta Cepstrum — SDC) [5]*

Частота дискретизации звуковых данных в обоих случаях составляла 8 кГц с разрядностью 16 бит. Для вычисления коэффициентов MFCC длительность окна анализа выбиралась равной 16 мс, окна анализа следовали друг за другом с шагом 8 мс. Размерность результирующего вектора наблюдений составляла 12.

Длительность окна для вычисления вектора SDC такая же, как и у MFCC. Вектора получены путём конкатенации семи компонент, каждая из которых представляет из себя разность сдвинутых во времени кепстральных коэффициентов для смежных блоков [5].

Вероятностные модели

Полученный в результате предварительной обработки речевых данных набор векторов наблюдения полагается выборкой некоторой генеральной совокупности независимых случайных величин. Плотность распределения этих величин можно описать смесью однотипных распределений. Задача построения вероятностной модели языка сводится к задаче разделения смеси, т.е. оценки параметров распределения каждой из компонент смеси при условии, что общее число компонент и их вид заранее известны.

Рассмотрим два типа вероятностных моделей — смесь нормальных распределений и смесь нормальных распределений авторегрессии. В качестве обобщённого подхода к оценке неизвестных параметров применялся стандартный EM — алгоритм [3]. С помощью максимизации функционала (логарифма правдоподобия):

$$p(\theta \mid x) = \ln \prod_X p(x \mid \theta) = \sum_X \ln\, p(x \mid \theta) \to \max$$

настраиваются параметры каждой из компонент смеси. Оптимизация проводится итерационно методом покоординатного спуска.

Для смеси нормальных распределений базовым объектом является гауссиан-функция, описывающая плотность распределения случайной величины для нормального закона. Метод разделения такой смеси не предполагает дополнительных знаний о способе представления речевых данных, поэтому вероятностная модель подобного типа была построена как для MFCC, так и для SDC. Настраиваемыми параметрами каждой из компонент являются: её априорная вероятность, математическое ожидание и дисперсия.

Смесь нормальных распределений авторегрессии имеет более сложную структуру, позволяющую учитывать временну̀ю зависимость входных данных. Базовой единицей этой модели является функция вида:

$$N_{AR}(x_t \mid \alpha_1,\ldots,\alpha_K,\eta,\Lambda) = \frac{1}{\sqrt{(2\pi)^d \mid \Lambda \mid}} \exp\{-\frac{1}{2}(x_t - \sum_{k=1}^{K}\alpha_k x_{t-k} - \eta)^T \Lambda^{-1}(x_t - \sum_{k=1}^{K}\alpha_k x_{t-k} - \eta)\}, \quad (1)$$

где

$x_t \in R^d$ — переменная, наблюдаемая в момент времени t;

$\alpha_1,\ldots,\alpha_K \in R^d$ — коэффициенты авторегрессии, K — глубина авторегрессии;

$\eta \in R^d$ — постоянная составляющая;

$\Lambda \in R^{d\times d}$ — дисперсия шума (матрица ковариации).

Заметим, что если записать авторегрессионую модель дискретного сигнала:

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_K x_{t-K}, \quad (2)$$

где $\varepsilon_t \sim N(\eta, \Lambda)$ — ошибка линейного предсказания (или ошибка авторегрессии), то функция вида (1) есть плотность распределения ошибки линейного предсказания с математическим ожиданием и ковариационной матрицей Λ[1].

Подобно тому, как это сделано для стандартного EM-алгоритма [3], рассмотрим смесь таких плотностей и зададимся вопросом об оценке параметров смеси.

Смесью нормальных распределений авторегрессии называется функция вида:

$$N(x_t \mid \theta) = \sum_{j=1}^{M} \beta_j N_{AR}^j(x_t, \quad (3)$$

причём $\sum_{j=1}^{M} \beta_j = 1$,

а $\theta^j = (\alpha_1^j, \ldots, \alpha_K^j, \eta^j, \Lambda^j)$ — набор параметров каждой из компонент смеси.

Для того чтобы оценить все неизвестные параметры в выражении (3), воспользуемся методом максимума правдоподобия. Как показано в [3], при помощи искусственного введения вспомогательного вектора «скрытых» переменных функция правдоподобия может быть приведена к виду:

$$Q(\Theta) = \sum_{j=1}^{M} \sum_{i=1}^{N} g_{ij} \log(\beta_j) + \sum_{j=1}^{M} \sum_{i=1}^{N} g_{ij} \log(N_{AR}^j(x_i)) \quad (4)$$

здесь $\Theta = (\beta_1, \ldots, \beta_M, \theta^1, \ldots, \theta^M)$, а g_{ij} — значение «скрытых» переменных, которые определяют, с какой вероятностью i-ый объект был сгенерирован именно j-ой компонентой. Имея некоторые априорные знания о начальном распределении параметров в смеси, эти значения несложно найти, используя формулу Байеса.

Первое слагаемое в выражении (4) не зависит от типа распределения каждой компоненты в смеси. Используя правило множителей Лагранжа и решая задачу условной оптимизации с условием $\sum_j \beta_j = 1$, найдём, что

$$\beta_{onm}^j = \frac{\sum_{i=1}^{N} g_{ij}}{N} \quad (5)$$

Второе слагаемое функции правдоподобия в выражении (4) является квадратичным функционалом по параметрам $\alpha_1^j, \ldots, \alpha_K^j, \eta^j$, $j = 1, \ldots, M$, поэтому необходимым и достаточным условием экстремальности является равенство нулю соответствующих производных. Покажем, что нахождение оптимальных параметров в этом случае эквивалентно нахождению решения системы линейных алгебраических уравнений.

[1] На практике вместо ковариационной матрицы используется вектор дисперсии в предположении о некоррелируемости случайных величин, что практически не ухудшает результат, но значительно упрощает вычисления.

Для любого $s = 1,\ldots,K$ имеем, что

$$\frac{\partial Q}{\partial \alpha_s^j} = \sum_{i=1}^{N} g_{ij}\left[-2\frac{1}{2}(-x_{i-s}^T)\Lambda_j^{-1}(x_i - \sum_{k=1}^{K}\alpha_k^j x_{i-k} - \eta^j)\right] = 0$$ — условие экстремальности.

Далее, ввиду того, что оптимизация параметров каждой из компонент может проходить независимо друг от друга, опустим индекс j.

Из предыдущего выражения следует:

$$\sum_{i=1}^{N} g_i x_{i-s}^T \eta + \sum_{k=1}^{K}\alpha_k \sum_{i=1}^{N} g_i x_{i-s}^T x_{i-k} = \sum_{i=1}^{N} g_i x_{i-s}^T x_i \quad s = 1,\ldots,K \qquad (6)$$

Это система из K уравнений с K+1 неизвестной. Как было отмечено выше, параметр η есть математическое ожидание ошибки линейного предсказания, которая распределена нормально. Поэтому, используя представление из [3], оптимальное значение параметра η имеет вид:

$$\eta = \frac{\sum_{t=1}^{N} g_t (x_t - \sum_{k=1}^{K}\alpha_k x_{t-k})}{\sum_{t=1}^{N} g_t} \qquad (7)$$

Подставляя указанное выше выражение в (6), получим

$$\frac{\sum_{i=1}^{N} g_i x_{i-s}^T}{\sum_{t=1}^{N} g_t}\left(\sum_{t=1}^{N} g_t x_t - \sum_{k=1}^{K}\alpha_k \sum_{t=1}^{N} g_t x_{t-k}\right) + \sum_{k=1}^{K}\alpha_k \sum_{i=1}^{N} g_i x_{i-s}^T x_{i-k} = \sum_{i=1}^{N} g_i x_{i-s}^T x_i$$

$$s = 1,\ldots,K$$

Для упрощения записи введём обозначения:

$$\sum_{\gamma} g_\gamma = r, \quad \sum_{\gamma} g_\gamma x_{\gamma-l} = r_l, \quad \sum_{\gamma} g_\gamma x_{\gamma-l_1}^T x_{\gamma-l_2} = r_{l_1,l_2},$$

и тогда система примет вид:

$$\frac{r_s}{r}\left(r_0 - \sum_{k=1}^{K}\alpha_k r_k\right) + \sum_{k=1}^{K}\alpha_k r_{s,k} = r_{s,0} \quad s = 1,\ldots,K$$

Домножая на r и перенося свободные слагаемые в правую часть, в итоге получаем

$$\sum_{k=1}^{K}\alpha_k \left(r_{s,k} r - r_s r_k\right) = r_{s,0} r - r_s r_0 \quad s = 1,\ldots,K \qquad (8)$$

То есть искомые коэффициенты авторегрессии являются решением системы линейных алгебраических уравнений (8) с симметрической матрицей.

Если известны коэффициенты регрессии (а следовательно, и ошибка), то можно найти оптимальные параметры так же, как это сделано в [3].

Итак, выпишем все формулы, дающие оценку «новым» параметрам через «старые»[2].

[2] «Старые» параметры входят в указанное выражение неявно, будучи используемыми при подсчёте значений вероятностей $\{g_{ij}\}$ через формулу Байеса.

$$\beta^{new} = \frac{\sum_{i=1}^{N} g_i}{N},$$

$$A\alpha^{new} = b,$$

$$\eta^{new} = \frac{\sum_{i=1}^{N} g_i (x_i - \sum_{k=1}^{K} \alpha_k^{new} x_{i-k})}{\sum_{t=1}^{N} g_i}$$

$$\Lambda^{new} = \frac{\sum_{i=1}^{N} g_i (x_i - \sum_{k=1}^{K} \alpha_k^{new} x_{i-k} - \eta^{new})(x_i - \sum_{k=1}^{K} \alpha_k^{new} x_{i-k} -$$

$$\sum_{t=1}^{N} g_i$$

Изложенный метод позволяет найти параметры, при которых достигается локальный максимум функции правдоподобия. Сходимость и точность решения во многом зависят от выбора начального приближения параметров. Конечно, можно выбрать начальные значения случайным образом, но на практике такой подход приводит к тому, что различные эксперименты (включающие в себя обучение) приводят к различным значениям точности системы. При использовании EM-алгоритма более пригодным оказывается следующий подход: сначала в выборке находятся k классов таких, что расстояние между объектами класса до его центра меньше, чем расстояние до центров всех прочих классов. Потом за начальное приближение математического ожидания берётся среднее значение каждого класса, а за дисперсию (в предположении о некоррелируемости случайных величин) — соответствующее ему среднеквадратичное отклонение. Построение соответствующих классов можно проводить на основе алгоритма k-средних (k-Means).

Для смеси нормальных распределений авторегрессии описанный выше метод организации начального приближения использовать неприемлемо — данные линейно связаны друг с другом. Однако если использовать вместо этих данных ошибку авторегрессии из представления (2), то подход, описанный выше, станет корректным. Остаётся вопрос — как найти начальные коэффициенты регрессии? Здесь можно предложить следующую эвристику: для начала также находим k классов в исходной выборке. Далее для объектов каждого класса независимо от других классов производим «нормализацию»: вычитаем их среднее значение. Предположим, что получена выборка временно̀го ряда, сформированная нормальным процессом авторегрессии с нулевым средним. Тогда коэффициенты этой авторегрессии будут искомыми коэффициентами. Для нахождения коэффициентов используем уравнения Юла-Уолкера (Yule-Walker) [4] , возникающие как следствие минимизации ошибки линейного предсказания.

Критерий остановки описанной выше итерационной процедуры может быть двух типов:

1. Близость по какой-либо метрике оптимизируемых параметров, например,

   при заданной точности δ критерием может являться $\left\| \eta_{new}^{j} - \eta_{old}^{j} \right\|^2 < \delta^2$

2. Стабилизация функции правдоподобия.

Как показали эксперименты, в отличие от первого критерия второй позволяет добиться требуемой точности решения за существенно меньшее число итераций.

Следует также отметить, что число компонент в смеси является структурным параметром [6], поэтому какие-то компоненты могут плохо описывать реальное распределение. В этом случае их априорная вероятность оказывается малой, и их необходимо удалить. Задача оценки оптимального числа компонент смеси для произвольной выборки некорректна.

Решающее правило

В результате обучения формируется s языковых моделей $\{P_1, P_2, \ldots, P_s\}$, каждая из которых характеризуется своим набором компонент смеси: $P_i \sim (p_1^i, p_2^i, \ldots, p_{M_i}^i)$. Язык, которому соответствует i-ая модель, считается распознанным, если правдоподобие i-ой модели максимально.

Логарифм функции правдоподобия имеет вид:

$$MLE_i = \log \prod_{k=1}^{N} P_i(x_k) = \sum_{k=1}^{N} \log P_i(x_k) = \sum_{k=1}^{N} \log \sum_{j=1}^{M_i} \alpha_j^i p_j^i(x_k) \quad i = 1, 2, \ldots, s$$

В этих обозначения решающее правило формально выглядит следующим:

$$i^* = \arg\max_{i=1,2,\ldots,s} MLE_i$$

Результаты

Ниже в таблице приведена точность распознавания пяти языков для трёх описанных выше языковых моделей. Каждая модель включает в себя 50 кластеров.

Все обучающие базы состоят из звуковых файлов с частотой оцифровки 8 КГц и разрядностью 16 бит; общий объём каждой базы — 500 Мб. Данные получены в различных каналах записи: микрофон, цифровая телефонная линия с μ-law кодированием, аналоговый телефон.

Название языка	MFCC	SDC	MFCC + AR
Арабский	95%	78%	91%
Английский	66%	23%	70%
Китайский	92%	54%	91%
Русский	80%	49%	83%
Турецкий	51%	73%	43%
Общая точность:	**76,8%**	**55,4%**	**75,6%**

Эксперименты показали, что авторегрессионная модель языка является менее чувствительной по отношению к каналу записи при незначительном падении точности распознавания, по сравнению с MFCC моделью. Несоответствия искажений, вносимых каналами, компенсируются динамическими свойствами акустических признаков речевого сигнала.

Точность идентификации на основе SDC модели существенно меньше точности, указанной в [5].

Заключение

В данной работе мы рассмотрели и дали сравнительный анализ трём типам акустических моделей речи, используемых для задачи автоматической идентификации языка: кепстральные мел-коэффициенты (MFCC), смещённый дельта кепстр (SDC) и авторегрессионная модель кепстральных коэффициентов (MFCC+AR). Параметры смеси гауссовских распределений, описывающей языковую модель, настраивались посредством обучения на базе размером 500 Мб. Результаты распознавания получены для моделей, состоящих из 50 кластеров. Увеличение числа кластеров ведёт к повышению точности, однако временные затраты на обучение при этом резко возрастают.

Эксперименты показали, что качество системы во многом зависит от однородности искажений всех полученных данных и ухудшается, если данные получены из различных каналов связи. Учёт динамических свойств речевых признаков позволяет повысить устойчивость к каналу при распознавании языка.

Литература

1. *Аграновский А.В., Зулкарнеев М.Ю., Леднов Д.А., Можаев О.Г.* Автоматическая идентификация языка// Искусственный интеллект, № 4, 2002, изд. НАН Украины, Донецк, 2002. С. 142–150.

2. *T. Schultz, Q. Jin, K. Laskowski, A. Tribble, and A. Waibel,* “Speaker, Accent, and Language Identification Using Multilingual Phone Strings”, HLT 2002, San Diego, California, March 2002.

3. *Jeff A. Bilmes* «A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models», International Computer Science Institute Berkeley CA, April 1998.

4. *Gidon Eshel* «The Yule Walker Equations for the AR Coefficients».

5. *Pedro A.* Torres-Carrasquillo «Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features», Proc. Int’l Conf. Spoken Language Processing, Denver, Sep. 2002.

6. *Моттль В.В., Мучник И.Б.* Скрытые Марковские модели в структурном анализе сигналов. М.: ФИЗМАТЛИТ, 1999.

Леднов Дмитрий Анатольевич —

кандидат технических наук, старший научный сотрудник, руководитель отдела речевых технологий ООО «Стел — Компьютерные Системы», окончил Казахский государственный университет, г. Алма-Ата, защитил кандидатскую диссертацию в Ростовском государственном университете, г.Ростов-на-Дону.

Иващенко Юрий Станиславович —

студент 6-го курса Московского института радиотехники, электроники и автоматики, сотрудник отдела речевых технологий ООО «Стел — Компьютерные Системы».

Любимов Николай Андреевич —

студент 5-го курса факультета вычислительной математики и кибернетики МГУ, сотрудник отдела речевых технологий ООО «Стел — Компьютерные Системы».

О распознавании фонем с помощью анализа речевого сигнала в частотной и временнòй областях.

Приложение к распознаванию синтаксически связных фраз

В.Ю. Шелепов,
доктор физико-математических наук

А.В. Ниценко, А.В. Жук, Д.С.Азаренко

Как известно, один из классических способов акустического описания фонем основан на использовании их формантной структуры — областей спектральных максимумов. В настоящей работе предлагается новый подход к этой проблеме, который сочетает в себе частотный анализ (обработку сигнала набором полосовых фильтров) с анализом профильтрованного сигнала во временнòй области.

Имея отрезок в 256 последовательных отсчётов записанного сигнала

$$x_1, x_2, ..., x_{256}$$

определим численный аналог полной вариации этого отрезка:

$$\sum_{i=1}^{256} \left| x_{i+1} - x_i \right| \qquad (1)$$

Если мы теперь выделим некоторый участок записанного речевого сигнала

$$x_1, x_2, ..., x_n, ... \qquad (2)$$

то его вариацией назовём среднее величин вида (1), вычисленных для последовательных окон в 256 отсчётов.

В работах [6], [9] подробно описан разработанный нами механизм автоматической сегментации речевого сигнала, т.е. разбиения его на участки, соответствующие отдельным фонемам с одновременным отнесением их к гласным, голосовым согласным, шипящим или паузообразным звукам. Основываясь на этом, мы выделяем участок записанного

речевого сигнала, соответствующий какой-либо одной фонеме. Затем обрабатываем сигнал полосовыми фильтрами с полосой пропускания от 0 до 200 Гц (получаем для него значения с плавающей точкой), умножаем профильтрованный сигнал на коэффициент 10 и вычисляем вариацию V_1 для выделенного участка профильтрованного сигнала. Затем обрабатываем сигнал фильтром с полосой пропускания от 200 до 400 Гц и вычисляем вариацию V_2 для выделенного участка 10-кратно увеличенного профильтрованного сигнала. Применяя таким же образом последовательно полосовые фильтры с полосой пропускания 200 Гц и заканчивая фильтром от 4800 Гц до 5000 Гц, получаем набор чисел

.

$V_1, V_2, ..., V_{25}$ Мы будем использовать также численный аналог полной вариации «с переменным верхним пределом» выделенного участка (2) записанного сигнала:

$$V(0) = 0, \qquad V(n) = \sum_{i=0}^{n-1} |x_{i+1} - x_i|$$

Определим также следующую величину. Пусть N_1 — максимальное число такое,

что $V(N_1) \le 255$. Далее полагаем

$$W(n) = V(n) \quad \text{при} \quad 0 \le n \le N_1,$$

$$W(N_1 + 1) = 0, \quad W(n) = \sum_{i=N_1+1}^{n-1} |x_{i+1} - x_i| \quad \text{при} \quad N_1 + 1 \le n \le N_2,$$

где N_2 — максимальное число такое, что и так далее. В результате возникает массив чисел.

Возьмём среднее этих чисел для выделенной части сигнала. Это среднее условимся называть «вариационной мерой» или просто «мерой» *M* выделенной части сигнала. Такая величина вычисляется для результатов фильтрации с упомянутыми выше полосами пропускания и получается набор чисел

$$M_1, M_2, ..., M_{25}$$

Наконец, вычисляются величины

$$Z_i = V_i / M_i, \quad i = 1,2,...,25. \qquad (3)$$

Мы исходим из представления о том, что фонема и слово — это акустически принципиально разные фонетические объекты. Фонема (и даже класс близких фонем) — объект спектрально сравнительно однородный, слово же, напротив, состоит из спектрально разнородных частей. Поэтому распознавая слова целиком, мы должны использовать тот или иной *вектор* признаков. Для распознавания же фонем (и их классов) более целесообразно использовать подходящий скалярный признак или набор независимых скалярных признаков, каждый из которых должен обеспечивать свой результат распознавания.

Речь — сложная система, которая в целом вполне детерминированна. В то же время общеизвестно, что любой мыслимый признак, который можно использовать при её распознавании, является случайной величиной. В этом нет противоречия,

подобные вещи происходят и в физике, когда при хаотическом движении отдельных молекул для их большого числа вырабатываются детерминированные макрохарактеристики, такие как температура, давление и т.д. Каждый признак следует рассматривать как случайную величину со своей функцией распределения, которая зависит от конкретного диктора, конкретного микрофона, конкретной звуковой карты. Последние два момента определяющие: какой смысл делать распознаватель инвариантным по отношению к диктору, если он зависит от микрофона? Мы считаем, что до тех пор, пока вопросы, связанные с независимостью от аппаратного обеспечения, не решены, целесообразнее разрабатывать быстро обучаемые системы распознавания речи с подстройкой под диктора. Аккуратное описание функции распределения — серьёзная задача, требующая большого статистического материала. В виду крайне ограниченной сферы применения (конкретный диктор, микрофон и т.д.) более или менее точное описание функции распределения становится нецелесообразным. В то же время, как правило, на основе нескольких примеров можно указать интервалы, куда чаще всего попадают значения признака для каждого члена рассматриваемой пары фонем. Значения за пределами этих интервалов разумно интерпретировать как отказ от распознавания.

Создавая обучаемую систему распознавания пары фонем, использующую один скалярный признак *X*, задаём два числа *a, b*. При

$$X < a \qquad (4)$$

считаем, что объект распознавания принадлежит первому классу, при

$$X > b \qquad (5)$$

— второму. При

$$a < X < b$$

не выполняется ни (4), ни (5), и мы имеем отказ от распознавания.

Вначале задаётся достаточно малое инициальное значение *a* и достаточно большое инициальное значение *b*. Если, пользуясь ими при распознавании, система не определит объект первого класса, то число *a* слишком мало. После того как пользователь укажет истинный результат, система должна заменить *a* вычисленным значением признака, увеличив последнее, скажем, на 0,1. Таким образом, в процессе обучения число *a* может только расти. Аналогично число *b* может только убывать. При этом обеспечивается все большая надёжность в случае принятия решения. Если, начиная с некоторого момента, окажется

$$a > b \qquad (6)$$

то при попадании *X* в *(b,a)* для распознаваемого объекта выполняются оба неравенства (4) и (5), т.е. он должен быть отнесён к обоим классам сразу, что невозможно, так как предполагается, что класс должен определяться однозначно. Таким образом, в случае (6) попадание *X* в *(b,a)* должно означать отказ от распознавания. Суммируя сказанное, получаем, что при

$$X < \min(a.b)$$

объект относится к первому классу, при

$$X > \max(a, b)$$

— ко второму классу, при

$$\min(a,b) < X < \max(a,b)$$

система отказывается от распознавания. Обучение состоит в модификации констант *a,b* и продолжается до тех пор, пока система не проработает без ошибок на протяжении, скажем, десяти циклов распознавания. Тогда распознаватель будет либо с высокой надёжностью принимать правильное решение, либо откажется от распознавания.

Теперь представим себе, что для полученной системы вероятность отказа от распознавания достаточно мала. Если мы для той же пары введём ещё несколько таких систем, использующих другие признаки, то, в соответствии со схемой независимых испытаний Бернулли, вероятность того, что все они одновременно будут отказываться от распознавания, станет существенно меньше. Все вместе построенные системы дадут желаемый распознаватель для рассматриваемой пары фонем, если в случае противоречия в результатах отдельных систем решение будет приниматься «по большинству голосов».

Запишем речевой сигнал, содержащий какую-либо пару гласных фонем, например, *А,И*, произнесённых как ударные (без редукции). Выделив участки сигнала, соответствующие этим фонемам, вычислим для них величины (3), которые обозначим через

$$Z_i(A) \quad \text{и} \quad Z_i(И), \qquad i = 1,2,...,25 \ . \qquad (7)$$

Для выделения компонент, лучше других различающих рассматриваемые фонемы, используем отношения величин (7). Пусть числа *k,l* таковы, что

$$Z_k(A)/Z_k(И) = \max_{1\le i\le 25}[Z_i(A)/Z_i(И)]$$

$$Z_l(A)/Z_l(И) = \min_{1\le i\le 25}[Z_i(A)/Z_i(И)] \ .$$

Тогда в качестве основного признака для различения *А, И* между собой возьмём величину

$$X(A,И) = Z_k / Z_l \ .$$

Настройка порогов для двухпорогового скалярного распознавателя описана выше. Для одного из авторов этой статьи и используемых им микрофона и звуковой карты при распознавании *А, И* между собой оказалось достаточно одного распознавателя описанного типа с параметрами *k=7,l=2, a=b=0,2*. Высокая надёжность распознавания в данном случае обеспечивается даже без заранее предусмотренного интервала отказа от распознавания.

Построим распознаватель такого типа для каждой пары следующего набора гласных:

А,Ё,И,О,У,Ы,W,Q (8)

Здесь символом *W* обозначено ударное *Е*, символом *Q* — ударное *Я*. Введение специальных обозначений для ударных *Е, Я* связано с тем, что только они имеют

достаточно определённое произношение. В безударном варианте они произносятся различными носителями языка по-разному. Для так называемой «младшей нормы» (более молодое поколение москвичей) они ближе к *И*, у сибиряков и в сценической речи — ближе к *Е*, *Я*. Для ряда пар не удаётся избежать случаев отказа от распознавания. В этом случае распознаватель выдаёт в качестве результата оба соответствующих символа. Если для какой-то пары вводятся дополнительные распознаватели, так, что общее число распознавателей для пары оказывается больше одного, то для них вычисляется совокупный результат «по большинству голосов». После всех попарных распознаваний результатом распознавания считается одна или несколько гласных, которые получились при распознаваниях пар максимальное число раз. Снабдив программу набором соответствующих флажков, мы получаем также возможность ограничивать распознавание лишь некоторыми гласными (8). Тогда остальные автоматически включаются в класс распознанных, поскольку на этом этапе для нас главное, чтобы слово «не потерялось», т.е. попало в формируемый программой список кандидатов на распознавание.

Таким же образом организуется распознавание в множестве голосовых согласных, множестве шипящих и аффрикат, множестве паузообразных звуков *К, П, Т, Ф, Х*. При распознавании в каждом из этих множеств в качестве результата в общем случае получается некоторый класс — часть этого множества фонем. Этот класс формируется автоматически. На *рисунке 1* представлен результат распознавания слова **«лиса».**

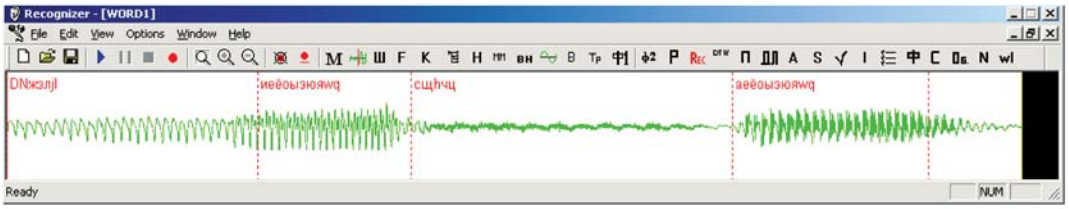


Рис. 1. Результат распознавания слова «лиса»

В работе [7] описан способ распознавания с предварительным заданием классов фонем, например, классов вида

аяqоёэеwуюыи W
аяqоёэеw Q
эewоёуюиыя Y
аяq A (9)
оёэеw R
эw E
оё O
.....

При этом распознаваемые цепочки из символов W,Q,Y,A,R… дают смешанную транскрипцию, состоящую из транскрипционных символов различного уровня детализации.

В идеале результатом распознавания фонем, образующих слово, служит его транскрипция, по которой слово в большинстве случаев однозначно восстанавливается. Однако любые признаки, используемые при распознавании речи, имеют характер случайных величин. Поэтому на любом этапе возможен отказ от распознавания и в результате вместо цепочки

транскрипционных знаков на выходе получается последовательность символов, обозначающих те или иные достаточно широкие классы фонем (смешение транскрипций разного уровня детализации). Возникает проблема, как по такому разнородному результату в большом словаре отыскать слова, которые ему удовлетворяют. В упомянутой работе [7] описан алгоритм, который основан на представлении транскрипций слов распознаваемого словаря в виде дерева и позволяет осуществить этот поиск очень быстро. В нашем случае распознанный класс фонем строится динамически. Введя для него виртуальное промежуточное обозначение типа (9), можно свести дело к алгоритму для заранее заданных классов (9). В результате получается способ поиска слов, соответствующих распознанной последовательности классов, причём длина поиска равна количеству распознанных классов, т.е. количеству фонем в сказанном слове. Это позволяет во времени, близком к реальному, находить списки слов-кандидатов на распознавание в случае словаря, содержащего миллионы словоформ.

Дальнейшее посвящено изложению концепции системы «речь — текст», позволяющей распознавать фразу как последовательность синтаксически связанных словоформ.

Известно, что в наиболее распространённом сейчас языке международного общения — английском — синтаксические связи между словами предложения осуществляются в основном с помощью предлогов. В результате за исключением некоторых моментов (типа «S» в конце слова в случае множественного числа), слово из текста всегда можно найти в орфографическом словаре. В отличие от английского русский язык относится к числу так называемых флективных языков. Большинство слов помимо начальной или словарной формы, называемой также «леммой», имеют достаточно развитую систему косвенных форм, образуемых с помощью флексий — частей, изменяемых при склонении, спряжении и т.д. Правильное использование этих форм — непременное условие синтаксически связной русской речи. Наличие многочисленных косвенных форм создаёт дополнительные трудности при компьютерном распознавании русской речи, ибо каждая из косвенных форм является для компьютера новым словом, в результате чего резко возрастает объём распознаваемого словаря. Различие же между отдельными формами часто сводится к безударным гласным в окончании, различать которые на сегодняшний день при обычной речи не представляется возможным. Последнее связано с редукцией упомянутых безударных гласных.

Имея дело с технической системой (хотя бы и относящейся к искусственному интеллекту), каковой является система компьютерного распознавания речи, мы вправе предложить пользователю соблюдать некоторые правила, относящиеся к самой речи. Например, можно на первых порах настаивать на подчёркнутой артикуляции, когда при слитном произнесении слова в нём слегка подчёркивается слоговая структура, так что каждая гласная становится как бы ударной.

Далее может быть предложена специальная архитектура системы, которая наряду с распознаванием речи использует элементы выбора из небольших словарей. Опишем одну из таких возможных систем. Распознаватель работает с первоначальным списком, содержащим все словоформы слов известного словаря Зализняка [10]. Получается достаточно полный словарь **всего русского язы-**

ка. Запись сказанного слова происходит при нажатии клавиши, соответствующей его первой букве, так что первая фонема при распознавании фактически заранее задаётся. В результате описанной выше процедуры пофонемного распознавания мы получаем список слов-кандидатов на распознавание — набор словоформ.

Далее используется наличие быстрого лемматизатора (система, восстанавливающая начальную форму слова по косвенной). Применяя к словам полученного списка лемматизатор, получим соответствующий набор начальных форм, который в несколько раз короче ранее полученного списка словоформ. По указанию пользователя (щелчок мыши на элементе списка) по исходному звуковому файлу строится эталон [5] и ему сопоставляется соответствующая лемма. Тогда, как показывает опыт, в подавляющем большинстве случаев при распознавании с использованием алгоритма DTW для любой словоформы этого слова построенный эталон окажется ближайшим и, следовательно, она будет отождествляться с указанной леммой. Исключения составляют ситуации типа «ИДТИ-ШЁЛ», «ЧЕЛОВЕК-ЛЮДИ». Далее компьютер отбирает из всех распознанных словоформ список словоформ, соответствующих распознанному слову и записанному сигналу. Это, собственно, и является результатом распознавания. В дальнейшем при произнесении других словоформ слова, для которого создан эталон, компьютер будет распознавать эти словоформы, используя эталон. Подчеркнём, что эталон строится по произвольной произнесённой словоформе, получает имя соответствующей леммы, и по нему распознаются (с точностью до леммы) все другие словоформы этого слова. Отметим также что, хотя дело заканчивается распознаванием по эталонам, применение пофонемного распознавания позволяет использовать эталоны в пределах списка распознанных словоформ, который на порядки меньше исходного полного словаря словоформ. На *рисунке 2* представлен результат распознавания слова **«солому»** по эталону, образованному по слову **«соломе»**.

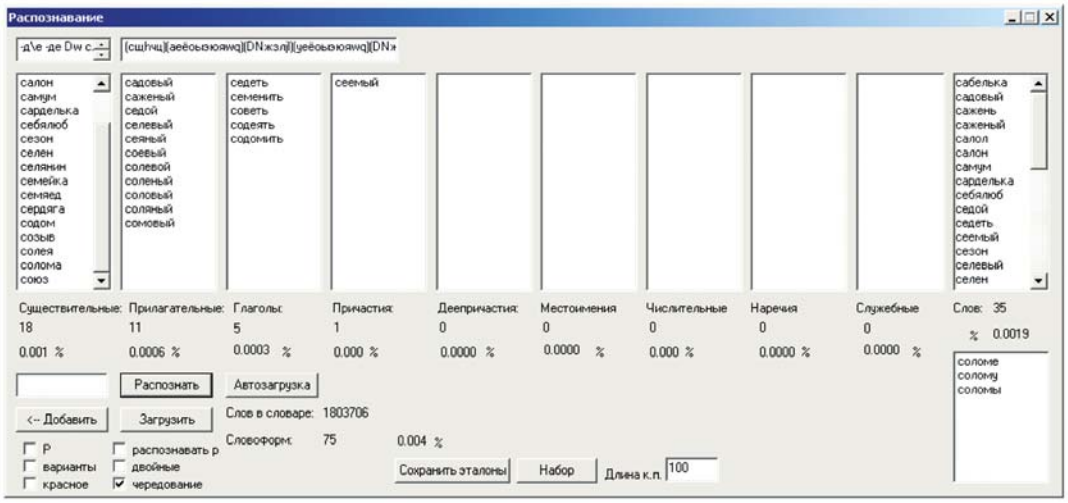


Рис. 2. Результат распознавания слова «солому» по эталону, образованному по слову «соломе»

В правом нижнем углу — список распознанных словоформ. При диктовке этот список передаётся в редактор, причём все его слова заключаются в скобки. Если список состоит только из одного слова, оно в скобки не заключается.

Пусть, например, диктуется фраза «Машина остановилась за углом». Тогда в окне редактора появятся следующие группы слов:

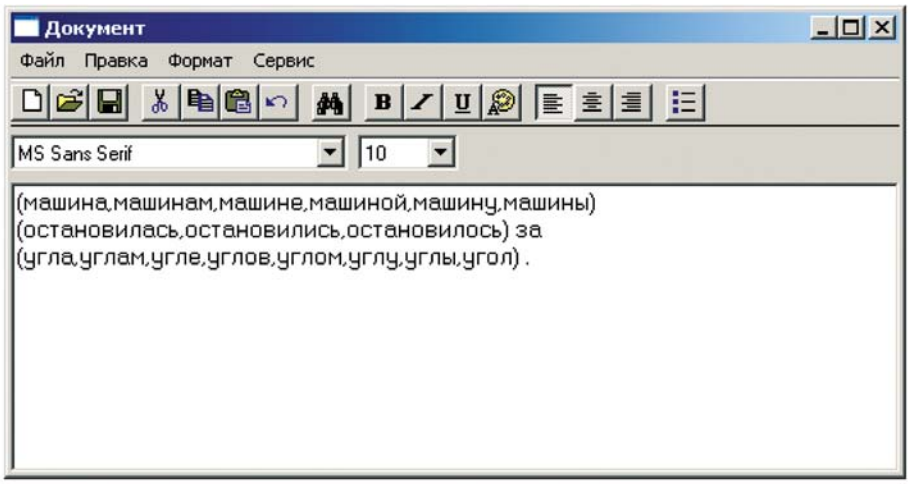


Рис. 3. Перечень групп при распознавании предложения

Далее, после того, как пользователь, закончив диктовать предложение, поставит точку, вступает в действие модуль синтаксической коррекции, работающий с использованием морфоанализатора, разработанного Г.В. Дорохиной [11, 12] и А.П. Павлюковой. Пользователь щелчком указывает подлежащее и выбирает для него нужную словоформу (в нашем примере — это слово «машина»). Компьютер убирает в отмеченной группе все лишние словоформы. Если при подлежащем есть прилагательное, в соответствующей группе автоматически оставляется только форма, согласующаяся с подлежащим. Аналогичным образом с подлежащим согласуется глагольное сказуемое. Среди форм существительного, следующего за предлогом, выбираются лишь те, которые этим предлогом допускаются. В нашем примере после первого шага получается следующее:

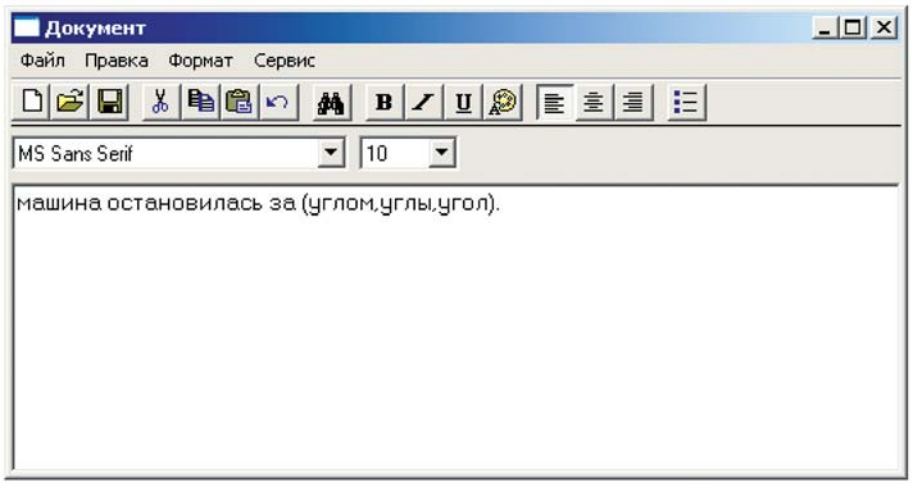


Рис. 4. Результат после выбора подлежащего и автоматического согласования с ним сказуемого

Конечный результат получается после выбора нужной словоформы в последней группе:

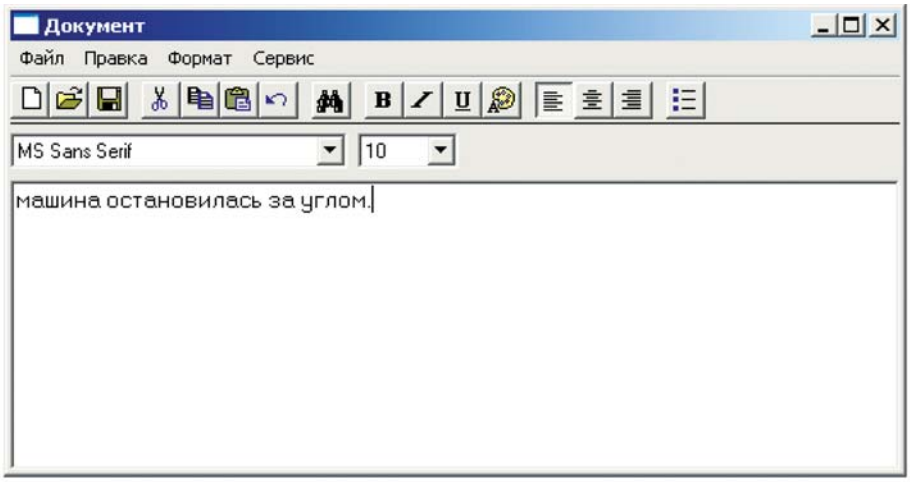


Рис 5. Конечный результат работы системы

Литература

1. *Дорохин О.А., Старушко Д.Г., Фёдоров Е.Е., Шелепов В.Ю.* Сегментация речевого сигнала // Искусственный интеллект. 2000. № 3. С. 450–458.

2. *Шелепов В.Ю., Ниценко А.В.* Амплитудная сегментация речевого сигнала, использующая фильтрацию и известный фонетический состав // Искусственный интеллект. 2003. № 3. С. 421–426.

3. *Ниценко А.В., Шелепов В.Ю.* Алгоритмы пофонемного распознавания слов наперёд заданного словаря // Искусственный интеллект. 2004. №. С. 633–639.

4. *Шелепов В.Ю., Ниценко А.В.* К проблеме пофонемного распознавания // Искусственный интеллект. 2005. № 4. С. 662–668.

5. *Засыпкин А.В., Мицевич А.Т., Овецкий М.В., Шелепов В.Ю.* О дикторонезависимой системе голосового телефонного номеронабирателя // Труды международной конференции «Знание-Диалог-Решение». Ялта. 1995. С. 427–430.

6. *Шелепов В.Ю., Ниценко А.В.* Структурная классификация слов русского языка. Новые алгоритмы сегментации речевого сигнала и распознавания некоторых классов фонем // Искусственный интеллект. 2007. № 1. С. 213–224.

7. *Шелепов В.Ю., Ниценко А.В., Жук А.В.* Новые алгоритмы распознавания фонем и их классов, поиск слова по его смешанной транскрипции при распознавании слов большого словаря // Искусственный интеллект. 2007. № 2. С. 139–147.

8. *Шелепов В.Ю., Ниценко А.В.* О распознавании фразы как последовательности синтаксически связанных словоформ // Искусственный интеллект. 2007. № 3. С. 344–346.

9. *Шелепов В.Ю.* Концепция пофонемного распознавания отдельно произносимых слов русской речи. Распознавание синтаксически связных фраз: Материалы международной научно-технической конференции Искусственный интеллект // Интеллектуальные системы (ИИ-2007). Донецк-Таганрог-Минск. С. 162–170.

10. *Зализняк А.А.* Грамматический словарь русского языка. М.: Русский язык. 1977.

11. Патент України № 78806 «Пристрій для збережёння і пошуку рядкових величин та спосіб збережёння і пошуку рядкових величин» Власник: Інститут проблем штучного інтелекту, винихідник Дорохіна Г.В. // Промислова власність. Бюл. № 5. 25.04.2007.

12. *Дорохина Г.В., Павлюкова А.П.* Модуль морфологического анализа слов русского языка // Искусственный интеллект. 2004. № 3. С. 636–642.

Шелепов Владислав Юрьевич —

*доктор физико-математических наук, профессор,
главный научный сотрудник Института искусственного интеллекта,
Начальник отдела распознавания речевых образов,
речью занимается с 1993 года.*

Ниценко Артём Владимирович —

младший научный сотрудник отдела распознавания речевых образов Института проблем искусственного интеллекта МОН и НАН Украины.

Жук Александр Викторович —

и.о. начальника отдела распознавания речевых образов Института проблем искусственного интеллекта МОН и НАН Украины.

Азаренко Дмитрий Сергеевич —

инженер отдела распознавания речевых образов Института проблем искусственного интеллекта МОН и НАН Украины.

Особенности современной русской звучащей речи

Л.В. Златоустова,
доктор филологических наук, профессор

В современном мире с его убыстрением темпа жизни увеличивается и темп речи. Если ещё 50 лет назад среднезвуковое время составляло около 90 мс, то сегодня, особенно у молодого поколения, оно составляет 50–60 мс. Тенденция к ускорению темпа прослеживается в ряде языков индоевропейской группы. Безусловно, живой язык, испытывая социальные изменения в процессе своего развития, ориентируется и использует собственные языковые средства.

В зависимости от строя языка, его принадлежности к той или иной языковой группе, способы, прежде всего, компрессии речи изменчивы. В рамках индоевропейских языков такими средствами являются сокращение числа фонем, видоизменение дифтонгов, вплоть до монофтонгов в отдельных позициях. Так, в современном французском языке фонологическая оппозиция широкого и узкого [e] практически не наблюдается, даже в замедленном, ораторском стиле (анализу подвергались речи Ш. Де Голля и Н. Саркози). В немецком языке сокращение нисходящей части дифтонга стало типичным. В ряде случаев дифтонг реализуется как монофтонг; в английском наблюдается сокращение употребления артиклей, хотя в данном случае необходимо учитывать жанровую принадлежность речевого произведения: например, в научной речи такое сокращение максимально. Известную роль в наблюдаемом процессе имеет и сознательное стремление в публичных выступлениях к демократизации речи, к усилению её воздействия на широкие массы слушателей.

В данной работе пойдёт речь преимущественно о тех изменениях в звучащей русской речи, которые сегодня не воспринимаются как отклонения от орфоэпической нормы и часто не замечаются слушателем. Тотальным явлением (не только для России) является урбанизация. Результат её — влияние диалектных фонем на литературное произношение, что характерно для суперсегментальной сферы речи и частично сегментальной, которая в большинстве случаев продиктована суперсегментными моделями речи, а также внутренними процессами развития языка.

Процесс падения редуцированных гласных, начавшийся в XII веке, продолжается особенно интенсивно с середины XX века и по сегодняшний день.

Сильная редукция гласных в среднерусских говорах отразилась на произношении москвичей нашего века. Этого явления не наблюдается в языке петербуржцев, ибо Санкт-Петербург испытывал и испытывает влияние севернорусских диалектов без редукции или со слабой редукцией. В связи с этим интересны интернет-тексты, где обсуждается вопрос о том, чьё произношение лучше — московское или петербуржское. Верна ли такая

постановка вопроса? Вряд ли. Но, несомненно, имеющееся различие представляет интерес при исследовании современной звучащей речи. С этой целью кратко остановимся на сопоставлении ритмической организации фонетического слова в сравниваемых произносительных вариантах. Для московского варианта типично следующее: гласный [а] первого предударного слога в позиции перед узким гласным часто превышает ударные по длительности, может быть равен ему, либо незначительно короче — коэффициент 1,15. Разумеется, в этом процессе имеют значение и такие факторы, как открытый или прикрытый первый предударных слог, количество слогов в слове, качество правого и левого окружения предударного и ударного гласных.

Иные значения слог имеет при узком предударном и широком ударном, где предударный всегда короче ударного: коэффициент 1,87–2.2. Вместе с тем, сказанное не меняет типа ритмической модели фонетических слов; однозначное восприятие даже псевдослов обеспечивается совокупностью правил, усвоенных и переведённых в полный автоматизм, так же как набор типичных моделей значимых фонетических слов или ритмических структур. Эти правила для русской речи сводятся к бессознательному принятию решения о месте ударения на основании речевого опыта, т.е. с детства носителем языка усваивается небольшая система моделей фонетических слов — в пределах 20, причём частотных ещё меньше; в случаях совпадения предударных и ударных по времени, а часто и по качественным характеристикам, принимается решение в пользу второго слога, т.к. в противном случае первый заударный был бы редуцирован как представитель позиции максимальной редукции. Разумеется, имеет большое значение и знание лексемы, её смысла. Хотя многократно проведённые эксперименты по восприятию псевдотекстов давали устойчивое значение по верному определению границ слов и месте ударения. В проведении эксперимента аудиторами были студенты-филологи 1 курса до прослушивания предмета «Русская фонетика». Записано в течение четырёх лет 115 аудиторов с общим результатом правильности опознавания 98%. Близкие результаты показали подготовленные аудиторы с той же инструкцией — прослушать текст, где слова внутри синтагм не были разделены паузами. Инструкция предписывала псевдословоформы разделить пробелами, как в орфографическом написании, поставить знак словесного ударения. Но второй эксперимент проводился при прослушивании псевдотекстов на болгарском, английском, немецком языках. Для указанных языков результаты были выше 90%.

Сказанное говорит об устойчивости структурированности фонетических слов.

Приводимый ниже график частотности употребления ритмических моделей слово показывает близость частотных моделей в группе индоевропейских языков со значительным отличием от них в языке алтайской группы.

График 1 показывает частотное распределение ритмических моделей фонетических слов, иначе ритмических структур (РС), где по вертикали указаны %, по горизонтали — РС. В числителе дано количество слогов, в знаменателе — место ударения.

Несомненно, частотность встречаемости типа РС зависит от жанра текста, его целевой направленности, спонтанности речепорождения или чтения сформированного написанного текста.

График 1.

Ритмическая структура фонетического слова достаточно устойчива, хотя и подвергается определённой деформации, в зависимости от степени эмоциональности произнесения, степени выраженности фразовых выделительных акцентов, частотности встречаемости той или иной лексемы и грамматической принадлежности, места в высказывании.

Необходимость обращения к структурированности диктуется определением слабых и сильных частей слов по отношению к реализации гласных. В современной речи типично произнесение глухих редуцированных гласных как следствие компрессии и практически утраты их фонологической значимости. Однако встаёт вопрос: почему они сохраняют себя в глухом варианте? С большой долей уверенности можно сказать — для сохранения ритмической модели слова. Например, в двусложных словах между глухими смычными согласными в научной публицистической речи частотно употребление глухого [ъ] в словах [ропът], [топът], [опът]. Причём глухой гласный имеет достаточно большую длительность: 50–60 мс, в зависимости от индивидуального темпа говорящего и объёма синтагмы.

Все приводимые рисунки являются широкополосными динамическими спектрограммами, выполненными в среде Windows, с помощью программы Speech Analyzer.

На *рис. 1* представлен первый заударный глухой гласный в слове «ропот» [ропът].

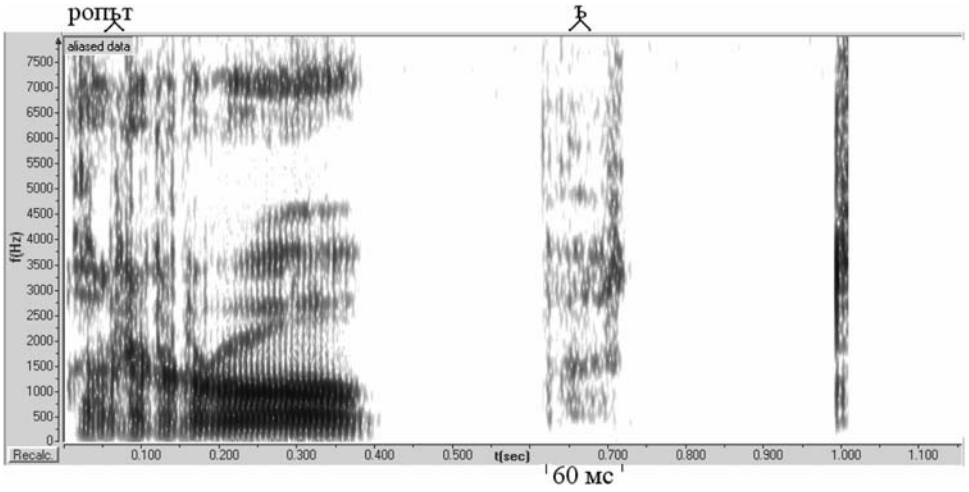


Рис. 1

Не менее интересны случаи «встраивания» глухих редуцированных звуков в согласные, прежде всего глухие щелевые, т.е. гласный реализуется одновременно с артикуляцией согласного. В приведённом ниже *рис. 2* глухой редуцированный гласный имеет три форманты: F2, F3, F4, но они выше нормативных значений.

На *рисунке 2* показан встроенный глухой гласный [ъ], а также конечный согласный глухой [γ].

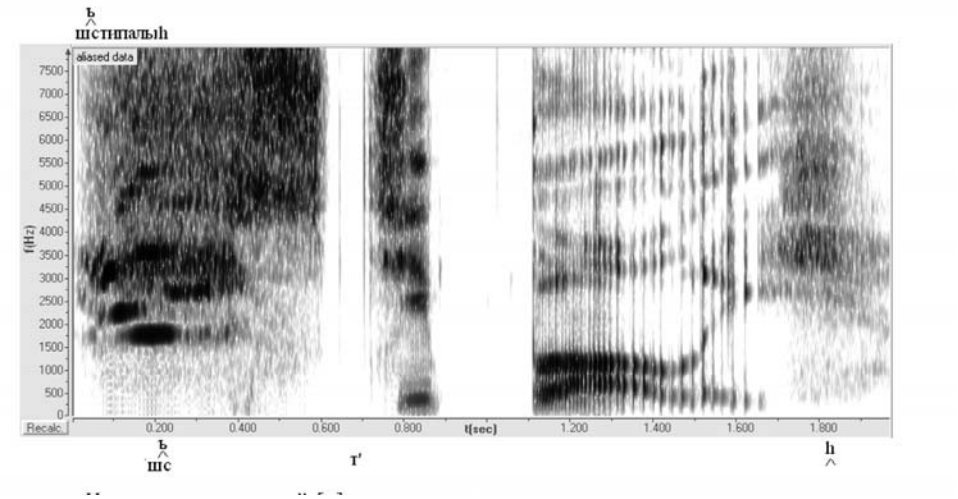


Рис. 2

Длительность «встроенного» гласного также больше по сравнению со звонким редуцированным. Спектрограмма слова [шъст'ипалъγ] включает частотный случай оглушения второго компонента дифтонга [ъi]. В концепции автора статьи все случаи реализации согласного [j], как гласного неслогового на основании спектрального анализа считаются элементами нисходящего дифтонга: под ударением [аi], [оi], [уi], в неударных позициях: [ъi], [ыi]. Например: майский маiский, моi, роi, дуi, красивыi, синиi и т.д.

Известно, что гласный [и] верхнего подъёма редуцируется, и примеров тому много. Но он редуцируется и в глухой вариант, в позиции между глухими щелевыми согласными, хотя имеются случаи редукции и между смычными глухими, например: [п'ьт'ир'им], [с'ьпаватыi] (Питирим, сиповатый).

Аналогично редуцируется этимологический [а]: [п'ьтачок], [п'ьт'ил'етка].
На *рис. 3* дана спектрограмма слова сиповатый [с' паватыi] с глухим редуцированным [и].

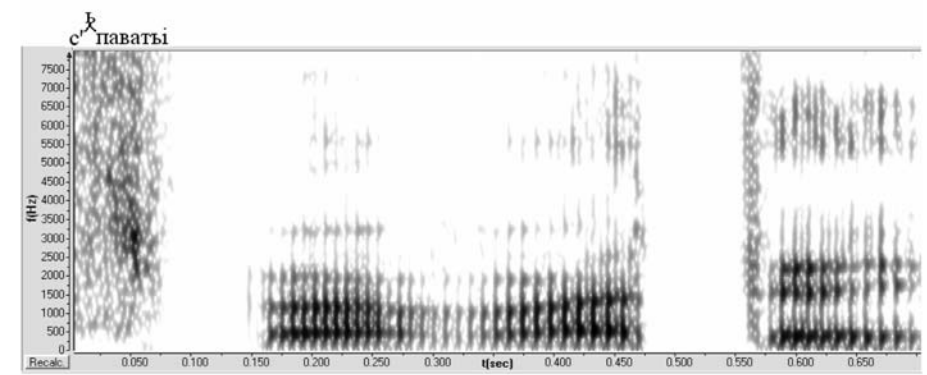


Рис. 3

На *рис. 4* показан «встроенный» редуцированный [ь] на месте этимологического [а], в позиции после мягких согласных.

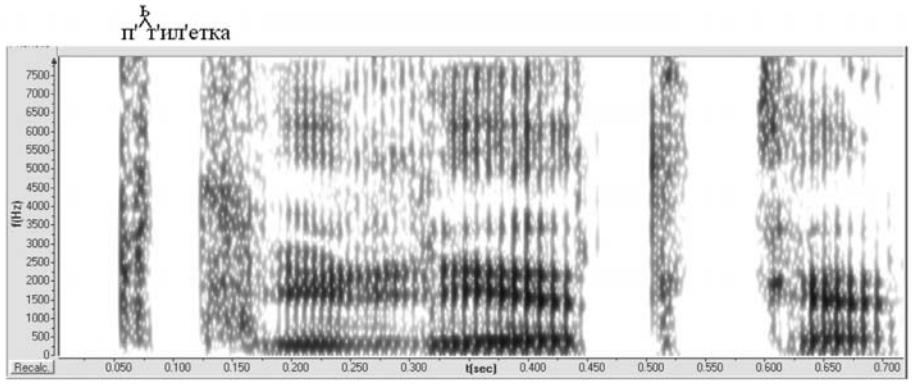


Рис. 4

Наиболее часты случаи реализации встроенных редуцированных, как и представленного на *рисунке 1,* в позиции между глухими щелевыми и аффрикатами. Характерно, что такого рода редукция охватывает не только гласные позиций второго, третьего предударных слогов, но и первого предударного, например: [ч'ьстотнъі], [с'ьстр'ич'ка], частотный, сестричка (*рис.5*).

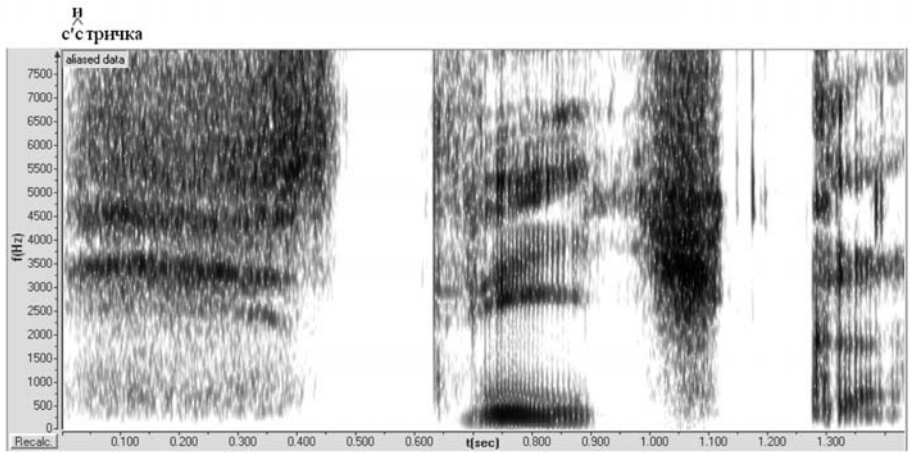


Рис. 5

Весьма типично употребление в разных жанрах квазиспонтанной речи усечения лексемы «человек» до формы [ч'ьк]. Такая усечённая лексема оказывается энклитикой, например: этот [ч'ьк] сказал, хороший [ч'ьк] был. Возможна произносительная форма [ч'ълк]. Реже у людей с нормативной речью в деловой и научной публичной речи встречается форма [ч'ьк] в начале высказывания ([ч'ьг] был). Кроме частотности употребления лексемы, человек имеет значение и формирование типичной ритмической модели слова ([ч'ьк] был — структура 2/2. В качестве примера приводим широкополосную спектрограмму *[этът ч'ьк] взял чек* (*рис. 6*).

Как показано на рисунке, при усечении слова [ч'ьк] редуцированный переднего ряда [ь] имеет три форманты, но при кратком времени, по сравнению со словом «*чек*», где [е] характеризуется как гласный полного образования.

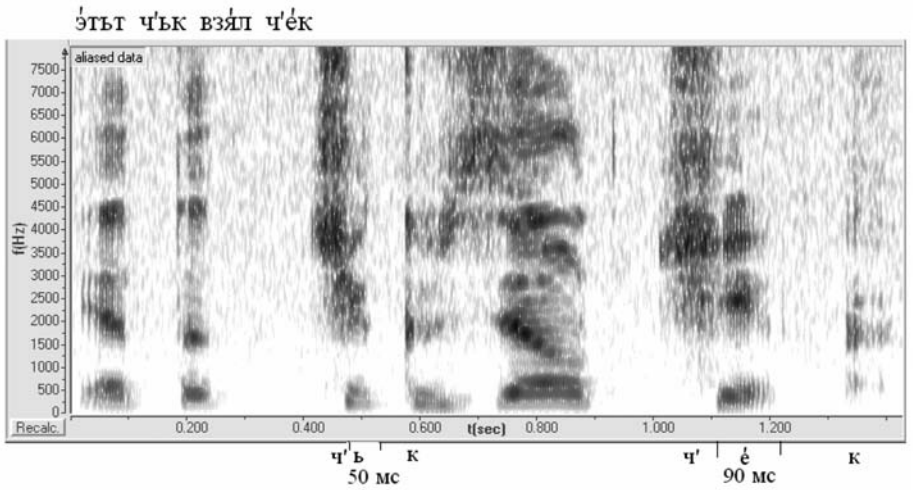


Рис. 6

По наблюдениям Л.В. Щербы и М.И. Матусевич, уже в середине XX века сонанты оглушались в позиции после глухих шумных согласных. Причём не всегда обращалось внимание на то, что в конце слов, стечения [пр', пл'] включают после глухих шумных согласных, перед сонантом и оглушённый гласный. В связи с анализом данной позиции встаёт вопрос о слогообразующей функции сонантов. Несмотря на спорность утверждения о том, что в рассматриваемой позиции дополнительный слог появляется только при наличии глухого гласного, и именно он образует слог, в нашем представлении обсуждаемого эффекта дело обстоит именно так. При отсутствии глухого гласного двусложность таких слов, как вепр', вопл' отсутствует, однако часто редуцированный имеет место.

На *рис. 7* представлено слово в'епър' как двусложное.

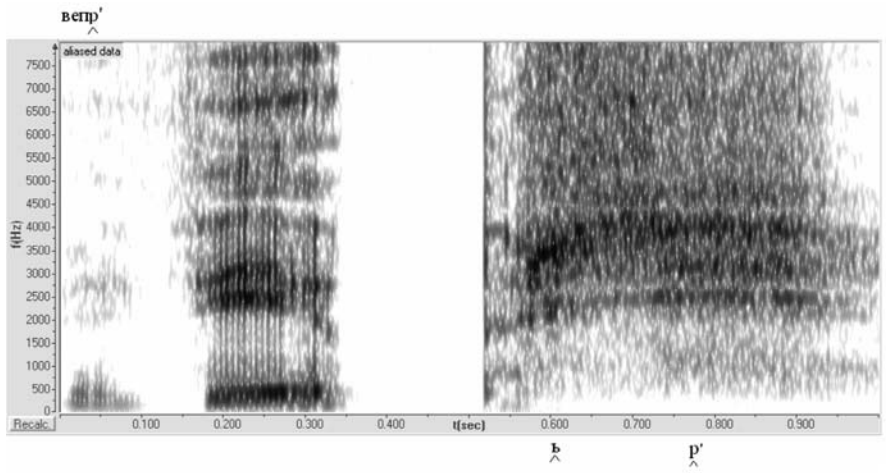


Рис. 7

Возвращаясь к случаям усечения (компрессирования) словоформ отметим частотность числительного *сколько*, произносимого как *скокъ*, или произнесение *ч'ьт* вместо *чёрт*, обычно оно наблюдается в относительно устойчивых словосочетаниях, как: *чёрт его знает [ч'ьтыво] знает*.

На *рисунке 8* дана спектрограмма рассматриваемых словоформ.

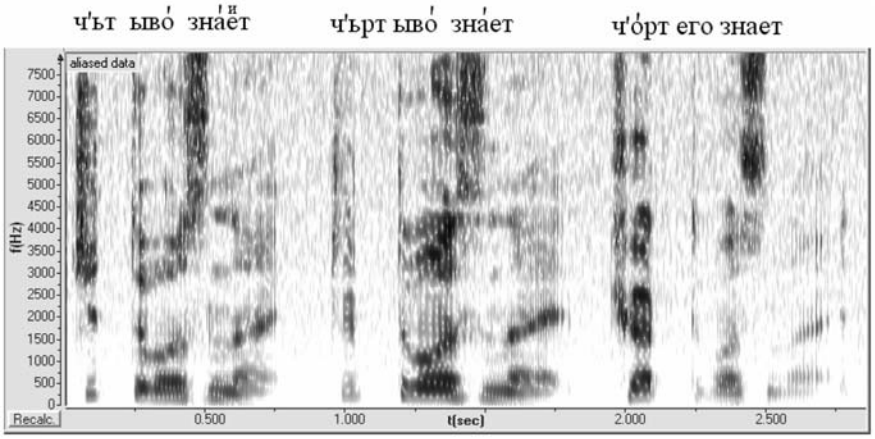


Рис. 8

Примером усечения гласных может служить форма эд:а вместо это да. Обычно реализовано в эмоционально окрашенном восклицании в типичной модели РС 2/2.

Современный процесс компрессии осуществляется параллельно с процессом развития русского языка в сторону аналитизма, что особенно заметно в таких процессах, как значительное увеличение словосложения как способа словообразования, унификация флексий, особенно имён прилагательных. Этот процесс весьма характерен в таких формах произнесения, как унификация падежей именительного и родительного в приводимых примерах: море сияет — мор['ь] сияет (*рис. 9*); моря сияющего нет (*рис. 9а*).

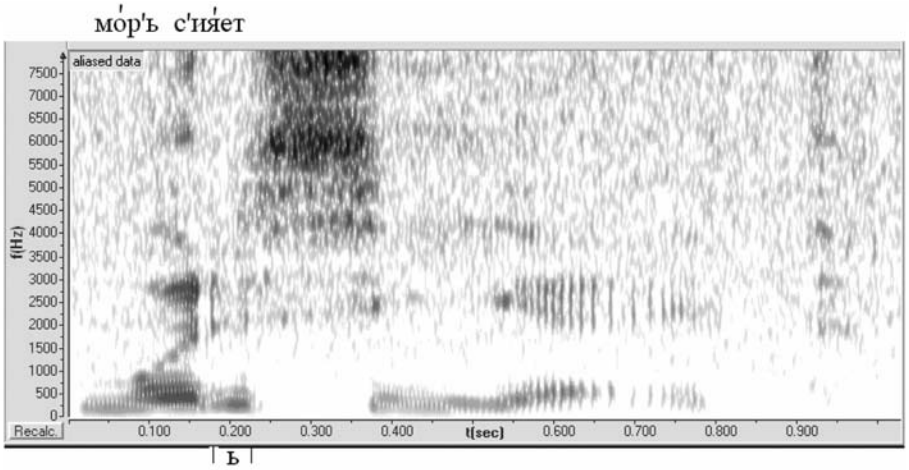


Рис. 9

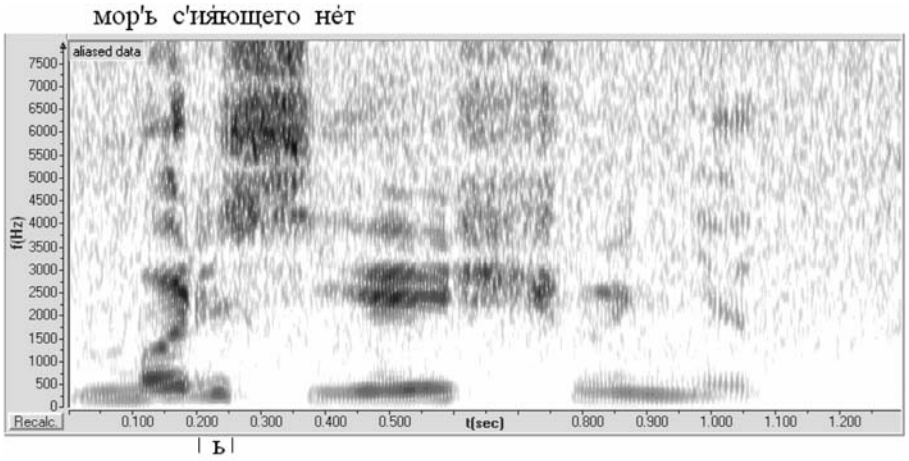


Рис. 9(а)

Унификация единственного и множественного числа в формах прилагательных: синяя лента — [с'ин'ьэь] лента, синие ленты — [с'ин'ьэь] ленты показана на рис.10.

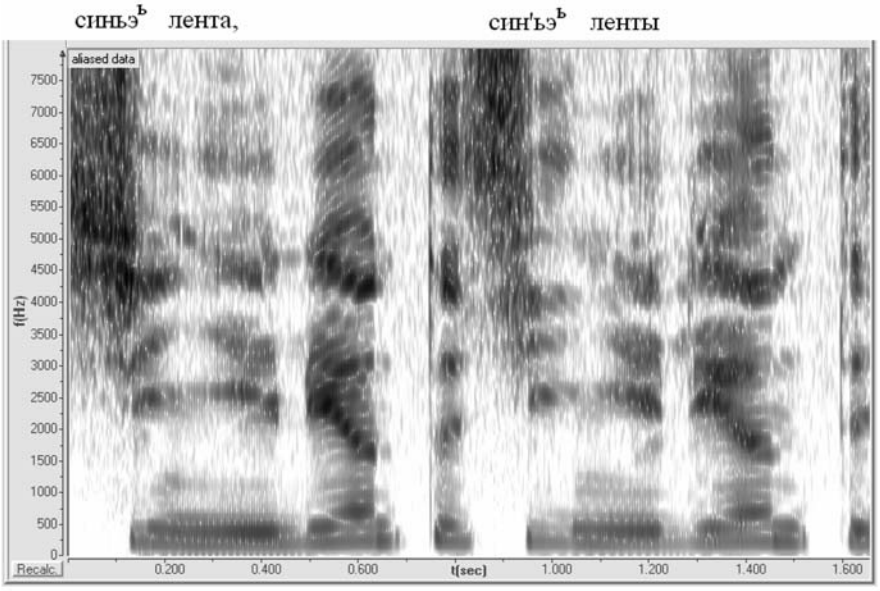


Рис. 10

Значительный интерес представляет изменение места ударения, что также часто обусловлено компрессией. Однако этот вопрос является предметом отдельного исследования, хотя и связан с настоящим.

Литература

1. Русская грамматика. Т.1. АН СССР. 1953 г.
2. *Zhlatoustova L.V.* Rhytmic Structure Types in Russian Speech. Auditory Analysis and Perception of Speech. 1975. London. New York. San Fransisco. P. 477–485.
3. *Златоустова Л.В.* «Наступление» диалектной фонетики на орфоэпическую норму русской речи. Сб. Фонетика сегодня. М., 2007. Стр. 67–70.
4. *Кривнова О.Ф.* Реализация словесного ударения в связном тексте. С. Фонетические чтения в честь 100-летия со дня рождения Л.Р. Зиндера. Санкт-Петербург, 2004. Стр. 150-155.
5. *Кодзасов С.В.* Комбинаторные методы в фонологии. Д/д. М., 2001.
6. *Зубкова Л.Г.* Фонологическая типология слова. М., 1990.
7. Horga Damir. Obrada fonetskih obavijesti. Zagreb, 1996.

Златоустова Любовь Владимировна

заслуженный профессор МГУ им. М.В. Ломоносова, член Международной комиссии по фонетике и фонологии, член президиума Российского акустического общества, действительный член Международной академии информатизации (президент отделения), действительный член Академии безопасности обороны и правопорядка.

Автоматическая расстановка огласовок в системах распознавания арабской речи

М.Ю. Зулкарнеев,
кандидат физико-математических наук

С.А. Репалов,
кандидат физико-математических наук

С.Х. Сальман,

О.А. Свирепо,
кандидат философских наук

The paper deals with automatic diacritization in systems of automatic speech recognition for the Arabic languages. The authors give an outline of the major difficulties in Arabic the researchers usually face when they start working on the problem of automatic diacritization. This part is followed by re review of methods and approaches that have been developed. The final part is devoted to the newly developed buckWalter algorithm which allows removing incorrectly diacritized words from the dictionaries.

Арабский язык занимает пятое место в мире по количеству говорящих на нём людей. Их число по разным источникам составляет 190–290 млн человек. Этот язык чрезвычайно привлекателен для разработчиков систем автоматического распознавания речи (САРР). Между тем количество работ, посвящённых САРР для арабского языка, невелико и их показатели значительно уступают САРР для других популярных языков.

Этому факту нетрудно дать объяснение. Дело в том, что арабский язык не имеет единой унифицированной формы. Само словосочетание «арабский язык» может подразумевать:

- классический арабский язык — язык Корана и сакральных текстов;
- так называемый современный стандартизированный арабский, или литературный арабский язык (Modern Standard Arabic), понятный большинству арабоязычного населения, это язык средств массовой информации, переговоров и т.д.;
- разговорный арабский, существующий в форме диалектов.

Классический арабский язык имеет узкую сферу применения, здесь мы встретим меньший набор лексических единиц при большом разнообразии грамматических форм, многие из которых не применяется в современной разговорной речи. Литературный арабский язык используется в основном для письменной коммуникации. Разговорный вариант языка реализуется в виде десятков разновидностей диалектов. Существенное расхождение между диалектами фиксировали ещё средневековые арабские источники, датируемые X веком.

В течение последующих столетий разница в фонетическом, лексическом и грамматическом плане только усиливалась. Именно разговорная форма является, с одной стороны, самой распространённой, с другой — характеризуется самой большой вариативностью на уровне фонетики, лексики и грамматики, что выражается в разнообразии диалектов: сирийский, египетский, ливанский, марокканский и т.д. Диалекты не имеют стандартизированной письменной формы и, следовательно, не зафиксированы в форме больших собраний записанных текстов. Сложности в создании текстовой базы данных привели к тому, что большинство работ по арабскому языку выполнено на материале литературного, реже классического варианта, в то время как большая часть живой разговорной речи реализуется в форме диалектов. Эксперименты, описанные в [1], показали, что современный стандартизированный арабский язык и диалектная форма (в данном случае использовался египетский диалект) ведут себя как два совершенно разных языка. В настоящее время для исследования диалектных форм арабского языка чаще всего используются базы данных LDC (Linguistic Data Consorcium) для левантийского арабского (57,3 часа речи, в том числе 3 часа транскрибированной речи) и «Call Home» (15 часов транскрибированной речи) для египетского диалекта. Небольшой объём данных затрудняет использовать статистические подходы. Наличие большой базы данных по арабским диалектам могло бы положительно сказаться на их точности. Согласно экспериментам [2] увеличение обучающей выборки в 10 раз улучшает модель языка на 3,5%, а увеличение выборки в 8 раз улучшает акустическую модель на 5%.

Диглоссия — не единственный фактор, осложняющий жизнь разработчикам САРР. Отличительная особенность арабской письменности — отсутствие графических средств для передачи кратких гласных. Исключение составляют учебные и сакральные тексты, где для обозначения кратких гласных используются специальные значки — огласовки. В результате одна и та же графическая единица, записанная несколькими согласными буквами, может иметь несколько вариантов прочтения.

О том, какое именно значение и какую фонетическую реализацию графической формы следует выбрать, носитель языка догадывается из контекста. Разницу между чтением на европейских и арабском языках отражает поговорка: «Европеец читает, чтобы понять, араб понимает, чтобы прочитать». Разговорные варианты арабского языка используют более богатый набор фонем, чем литературный язык: в диалектах часто встречаются звуки — e, -o, однако специальных графических средств для передачи этих звуков не предусмотрено. Работа с текстами без огласовок затрудняет построение языковых и акустических моделей. Сложно обучить акустическую модель, если краткие гласные нельзя идентифицировать внутри сигнала, а также точно узнать их положение. Модель языка, обученная на неогласованных текстах, имеет меньший предсказательный потенциал, чем модель, обученная на огласованных текстах. Оба этих фактора могут отрицательно влиять на точность распознавания.

Наконец третий фактор — сложность морфологии арабского языка, затрудняющая построение модели языка. Наличие большого количества суффиксов, аффиксов, моделей словообразования даёт такое количество словоформ, что для оценки модели языка требуются текстовые базы гораздо большего объёма, чем для английского. Одно слово может представлять собой целое предложение, например burrida — «он был охлаждён». Развитая морфология также приводит к ситуации, когда система часто сталкивается со словами, которых нет в словаре. Об актуальности этой проблемы свидетельствует то, что, согласно зарубежным источникам, если использовать все сценарии словообразования для всех слов арабского языка, его словарный состав включал бы 6 x 1010 единиц [3]. Это свойство арабского языка делает

проблематичным применение получившего широкое распространение статистического подхода — слишком часто встречаются слова, которых нет в словаре. По данным [1] около 10% слов в текстах, на которых проводилось испытание, не были найдены в словаре. Примерно 50% всех этих слов были всего лишь морфологическими вариантами словарных единиц, остальные 10% — именами собственными и 40% — действительно новые слова.

Подведём итоги: в чём специфика арабского языка для систем автоматического распознавания речи? Некоторые задачи распознавания решаются проще, чем в других языках, например, создание словаря произношений, поскольку арабский даёт практически однозначное соотношение буква — звук. Наибольшую сложность при разработке высокоточной системы автоматического распознавания речи представляет текстовый материал без огласовок, сильная вариативность диалектов и морфологические трудности.

Следовательно, автоматическая расстановка огласовок — одна из важных задач при разработке системы распознания арабской речи. Автоматическая расстановка огласовок предполагает морфологический разбор слова. На данном этапе исследований существует несколько возможностей для решения этой проблемы.

Морфологический анализатор Multi-Mode Morphological Processor (MMMP), предлагаемый компанией Sakhr Software на коммерческой основе. Программа позволяет идентифицировать все возможные основы слова, часть речи, произвести морфологический разбор. Кроме того, программа может проделать и обратную работу: составить слово из его морфологических частей (основа, корень, шаблон, часть речи, аффиксы). Предполагается, что программа даёт одинаково хорошие результаты как для современного арабского языка, так и для классических текстов. В научных исследованиях редко используют ПО Sakhr ввиду его высокой стоимости и невозможности получить свободный доступ хотя бы к некоторым возможностям программы. Кроме того, компания не раскрывает информацию, касающуюся алгоритмов и принципов работы морфологического анализатора.

Ещё один вариант морфологического анализатора представлен в [4]. Создатели этой версии использовали скрытые Марковские модели (СММ), обученные на вокализованных арабских текстах. Система обучалась на текстах Корана. Полученный результат — 85% правильно расставленных огласовок, но только для текстов Корана.

В [5] для решения задачи морфологического разбора исследователи использовали конечный автомат со взвешенными состояниями, обученный на базе данных LDC (Linguistic Data Consorcium). Заявленная точность составила 93% правильно огласованных текстов.

Два последних подхода имеют некоторые недостатки. В основе статистических подходов лежит предположение о том, что текст представляет собой последовательность наблюдений (Скрытая Марковская модель), где скрытыми состояниями являются возможные символы огласовок. В обоих случаях разработчики пошли по пути создания модели и последующего обучения её на материале корпуса текстов. Следовательно, обе модели зависят от корпуса текстов, а, например, коранический корпус не отражает современных реалий арабского языка. Помимо скрытых марковских моделей, возможно применение статистических методов, как это было сделано, например в [6]. Такой подход предполагает следующую последовательность действий:

- синтаксический разбор текста: текст делится на фразы, а фразы на слова;
- морфологический анализ текста: для каждого слова предлагаются все возможные правильные варианты огласовок;
- идентификация по частям речи — выбор правильного варианта огласовки зависит, в том числе, от того, к какой части речи относится данное слово;
- применение лингвистических эвристических правил, установленных лингвистами-экспретами с целью устранить возможные оставшиеся неправильные варианты.

Исследования, проведённые Safadi et al, не были основаны на достаточном количестве текстов, поэтому полученный результат 80–90% нельзя считать абсолютно достоверным.

Как было отмечено ранее, арабский язык отличается сложной морфологией, большим количеством словообразовательных моделей, что в сочетании с малыми обучающими выборками даёт проблему большого количества слов, отсутствующих в словаре. Поэтому применение чисто статистических методов, когда неогласованное слово заменяется огласованным вариантом, наиболее часто встречающимся в тексте, даёт неудовлетворительные результаты. Значительная доля ошибок приходится на неправильно огласованные окончания.

Нами был разработан алгоритм автоматической расстановки огласовок, основанный на шаблонах слов и списках корней. Для выполнения морфологического анализа мы использовали морфологический анализатор buckWalter.

Принцип состоит в следующем. Всего в арабском языке три части речи: имя, глагол и частица. Имена и глаголы образуются на основе 3–4 буквенных корней, реже встречаются 5-буквенные корни. Таким образом, гласные в арабском языке не являются полноправными элементами корня, а передают в основном грамматическую и словообразовательную информацию, поэтому такой текст читается носителями языка легче, чем текст на индоевропейских языках с пропущенными гласными. С этой точки зрения, приблизительным аналогом консонантного письма мог бы стать текст на индоевропейском языке, использующий сокращённую запись слов без окончаний и суффиксов.

Словообразование в арабском языке происходит при помощи шаблонов. Образование однокоренных слов, относящихся к другим частям речи или другими грамматическим категориям, изображают в виде шаблонов: согласные остаются неизменными, схематически показывается чередование и/или выпадение гласных. Число таких шаблонов весьма велико, поэтому создание программы автоматического морфологического разбора для арабского языка — чрезвычайно трудоёмкое занятие, например, для создания программы MORPHO3 потребовалось 3 человеко/года.

Морфологический анализатор buckWalter, помимо того, что доступ к этой программе при условии её некоммерческого использования свободный, показал себя как эффективная основа для реализаций автоматической расстановки огласовок. На вход морфологического анализатора подаётся некоторая графическая форма слова в виде последовательности согласных букв, на выходе мы получаем все возможные правильные варианты огласовок (фактическое произношение слова). Например:

عlm	@` allma
عlm	@` ulima
عlm	@` allama
عlm	@` ilm
عlm	@` alam

Морфологический анализатор включает в себя три словаря: префиксов (a), основ (b) и суффиксов (c). Предполагается, что

- длина префикса 0–4 символа;
- основа состоит из 1–10 символов;
- суффикс может иметь 0–6 символов.

В словарях содержатся все возможные правильные варианты произнесения префиксов, основ и суффиксов соответственно. Кроме того, элементам словаря присвоено определённое грамматическое значение. Помимо словарей, морфологический анализатор также включает три файла связей ab, bc, ac, где содержится информация о всех возможных правильных сочетаниях элементов всех трёх словарей. Таким образом, графическая форма анализируется на предмет наличия в ней трёх частей — суффикса, основы, префикса и их возможных грамматических значений. Окончательный вывод о всех возможных вариантах произнесения этой последовательности букв делается на основе анализа сочетаемости. Данная модель была внедрена, при этом были дополнены словари префиксов и суффиксов морфологического анализатора buckWalter. Очевидно, что этот способ автоматической расстановки огласовок может быть использован и для автоматического распознавания записей на различных диалектах арабского языка. В этом случае все три словаря морфологического анализатора необходимо дополнить соответствующими диалектными формами.

В результате работы мы получили быстродействующий алгоритм, используемый при составлении словарей распознавания, а также возможность замены и дополнения данных словарей за короткие промежутки времени, при этом структура алгоритма на основе правил арабского языка — как литературного, так и разговорного — позволяет исключить возможность вхождения некорректных форм огласования в результирующие словари.

Литература

1. *K. Kirchhoff et. al.,* Novel Speech Recognition Models for Arabic, Final Report, JHU Summer Research Workshop, Baltimore, MD, 2002.
2. *G. Zavaliagkos, J. McDonough, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Siu, H. Gish* «The BBN Byblos 1997 Large Vocabulary Conversational Speech Recognition System».
3. *Ahmad, Mohamed Attia* «A large-Scale Computational Processor of the Arabic Morphology, and Applications» A master's Thesis, Faculty of Engineering, Cairo University, Cairo, Egypt.
4. *Ya'akov Gal,* «An HMM Approach to Vowel Restoration in Arabic and Hebrew», ACL 02 Semitic Language Workshop, 2002.
5. *R. Nelken* ***и*** *S. Shieber* (Rani Nelken and Stuart M. Shieber, «Arabic Diacritization Using Weighted Finite-State Transducers», Proceedings of the 2005 ACL Workshop on Computational Approaches to Semitic Languages, pages 79–86, Ann Arbor, Michigan, June 2005.
6. *Hani Safadi, Dr. Oumayama Dakkak, Dr. Nada Ghnaim* «Computational Methods to Vocalize Arabic Texts», Proceed. of the 2006 Workshop on Internationalizing W3C's Speech Synthesis Makeup Languages.

Зулкарнеев М.Ю. —
кандидат физ.-мат. наук, старший научный сотрудник лаборатории обработки речи НИИ «Спецвузавтоматика», г. Ростов-на-Дону asni@asni.rsu.ru

Репалов С.А. —
кандидат физ.-мат. наук, заведующий лабораторией обработки речи НИИ «Спецвузавтоматика».

Сальман С.Х. —
научный сотрудник лаборатории обработки речи НИИ «Спецвузавтоматика».

Свирепо О.А. —
кандидат философских наук, старший научный сотрудник лаборатории обработки речи НИИ «Спецвузавтоматика.

Актуальные задачи речевой акустики

Е.М. Максимов,
доктор технических наук

Ю.Н. Ромашкин,
кандидат технических наук

С.А. Лопатина,
кандидат физико-математических наук

В статье даётся оценка достигнутого уровня развития отечественных речевых технологий по основным прикладным задачам. Указываются существующие ограничения, обусловленные сложившимися представлениями о процессах речеобразования и слухового восприятия речи. Формулируются приоритетные научные направления по дальнейшему совершенствованию речевых технологий.

Современный уровень развития методов и средств цифровой обработки сигналов предоставляет широкие возможности для внедрения речевых технологий в различные сферы жизнедеятельности, связанные с речевой коммуникацией. Практический интерес к речевым технологиям обусловлен как коммерческими потребностями, так и необходимостью решения специальных задач, возникающих, например, в речевой криминалистике [1].

Данная статья посвящена анализу существующих методов и алгоритмов цифровой обработки речевых сигналов, а также постановке задач, которые, по нашему мнению, являются актуальными с точки зрения повышения эффективности обработки в соответствующих этим методам практических приложениях.

Анализ современного состояния и направлений развития речевых технологий рассмотрим в рамках следующей классификации прикладных задач, решаемых с помощью этих технологий:

- идентификация (верификация) диктора по устной речи;
- распознавание естественной речи (в отличие от распознавания команд или изолированных слов), в частности, распознавание «ключевых» слов в слитной речи;
- повышение разборчивости речи на фоне акустических помех и искажений;
- синтез естественной речи.

К настоящему времени методы решения перечисленных задач основаны на сложившихся представлениях и моделях речеобразования и слухового восприятия. Эти модели с определёнными ограничениями и упрощениями использованы

при статистическом синтезе существующих алгоритмов обработки и позволили достичь известных положительных результатов. Вместе с тем, достигнутый уровень решения задач с помощью этих методов является недостаточным, а в ряде случаев неудовлетворительным для их практического применения. Для достижения качественно нового уровня необходимы, с одной стороны, разработка методов и алгоритмов, адекватных механизмам речеобразования и восприятия, с другой стороны — исследования, направленные на уточнение и более глубокое изучение этих механизмов.

Рассмотрим в данном контексте уровень развитых речевых технологий, применяемых в сформулированных выше прикладных задачах.

Автоматическая идентификация (верификация) диктора

Доминирующие позиции при решении данной задачи в настоящее время занимает GMM-метод [2], основанный на многомерной гауссовской аппроксимации выборочных плотностей распределения ряда акустических параметров речи. В приложениях, реализующих этот метод, используется либо неадаптивная GMM-модель небольшой размерности (24–64), либо адаптивная с высокой размерностью 1024–2048. В качестве информативных признаков, характеризующих индивидуальные особенности речи диктора, наиболее часто используются мел-кепстральные коэффициенты, их первые и/или вторые производные по времени. Такой подход, по нашим оценкам, позволил в среднем обеспечить следующие значения основных показателей эффективности работы алгоритмов:

- область работоспособности по отношению сигнал/шум — не ниже 10–15 дБ;
- минимальная длительность речи для составления модели голоса диктора — порядка 1 мин.;
- минимальная длительность речи для идентификации –10–15 с;
- вероятность правильной идентификации — 0,85–0,95;
- вероятность ложной тревоги — не более 0,10–0,15.

Дальнейшее повышение эффективности алгоритмов автоматической идентификации (верификации) возможно, по нашему мнению, по двум направлениям. С одной стороны, требуется совершенствование существующего GMM-метода в части обеспечения его инвариантности к преобразованиям и искажениям речевого сигнала в канале передачи и приёма, свойствам акустических помех и шумов канала, расширения вектора используемых акустических параметров речи и оптимизации размерности пространства информативных признаков. Более перспективным и актуальным, однако, представляется добавление в алгоритм идентификации ряда лингвистических признаков, характеризующих индивидуальные особенности порождения и структуру речи каждого диктора. Необходим поиск таких лингвистических признаков, способов их математической формализации и оценки с помощью цифровых методов анализа речевого сигнала.

Автоматическое распознавание естественной речи

Задача автоматического распознавания естественной речи (включая распознавание «ключевых» слов в слитной речи) является наиболее сложной, поскольку в наибольшей степени требует адекватных моделей порождения, восприятия и понимания речи. Исследованиями различных аспектов проблемы распознавания речи занимаются многие зарубежные и отечественные организации на протяжении нескольких десятилетий. Но, несмотря

на значительные успехи лабораторных разработок в этой области, практическое применение такие системы нашли в очень узкой области.

Современные подходы к решению данной задачи широко используют аппарат скрытых марковских моделей, параметрическое представление акустических параметров речи для различных элементов слитной речи (аллофонов, дифонов, трифонов и т.д. [3]). Далее осуществляется поиск наиболее вероятной последовательности распознанных звуков или слов с учётом принятой модели конкретного языка.

Принятый подход позволил, по нашим оценкам, в среднем обеспечить следующие значения показателей эффективности работы алгоритмов:

- область работоспособности по отношению сигнал/шум — не ниже 15–20 дБ;
- вероятность правильного распознавания «ключевых» слов в слитной речи — примерно 0,8–0,9 (при объёме рабочего словаря порядка 100 слов и вероятности ложной тревоги — не более 0,2–0,3);
- вероятность правильного распознавания слитной речи — около 0,6–0,7 (при объёме рабочего словаря порядка 20–30 тысяч слов).

Надёжность существующих систем автоматического распознавания речи зависит от многих факторов.

Речь, состоящая из изолированных слов или произносимая в замедленном темпе, более проста для распознавания. В быстрой естественной речи некоторые фонемы «смазываются» или просто «проглатываются», возрастают также коартикуляционные эффекты.

Наличие акустически похожих слов также затрудняет распознавание. Системы, рассчитанные на большой словарь, требуют больше времени на принятие решения. Уменьшение этого времени обычно производится за счёт упрощения алгоритма, что приводит к увеличению ошибок.

Сложность звукового строя конкретного языка в значительной степени определяется его фонетическим составом и правилами порождения слов. Например, звуковой строй японского языка гораздо проще для распознавания, чем французского, а русского и английского — сложнее французского.

Существенно влияют на надёжность автоматического распознавания речи воздействие внешних акустических помех, наличие амплитудно-частотных и временных искажений речевого сигнала в канале приёма и передачи, изменения психофизического состояния говорящего, артикуляционные дефекты в речи. Явление интерференции (смешение языков, взаимовлияние языков) давно интересует исследователей, но только сейчас делаются попытки их формализации. Необходима дальнейшая систематизация фонетических, лексических, синтаксических, просодических отклонений в речи иностранцев.

Без решения всего комплекса этих проблем получение качественно новых результатов при автоматическом распознавании речи представляется маловероятным.

Повышение разборчивости на фоне акустических помех и искажений

Эффективность алгоритмов выделения речи на фоне помех в значительной мере определяется акустическими условиями приёма речи и статистическими свойствами помех. Задачу подавления квазистационарных коррелированных помех и повышения разборчивости речи на фоне таких помех можно считать решённой как в теоретическом, так и прикладном плане на основе применения алгоритмов адаптивной фильтрации [4]. При наличии некоррелированных помех разработаны методы, основанные на теории марковской нелинейной фильтрации [5]. Однако предположения и допущения, используемые при реализации алгоритмов марковской фильтрации, не в полной мере адекватны механизмам речеобразования и восприятия, что обусловливает искажения речевого сигнала, хотя отношение сигнал/шум может быть существенно повышено. Эти же недостатки свойственны методам спектрального вычитания. Они позволяют увеличивать отношение сигнал/шум, однако, разборчивость речи при этом либо совсем не повышается, либо даже снижается вследствие появляющихся в результате обработки заметных искажений речи. Для условий приёма речи на фоне нестационарных помех (музыкальных, речевых и т.п.) пригодных для практического применения алгоритмов фильтрации пока не разработано.

В целом существующие методы выделения речи на фоне аддитивных помех умеренной интенсивности (отношениях сигнал/шум около 10 дБ) обеспечивают, по результатам наших оценок, следующие выигрыши в слоговой разборчивости речи:

- при наличии квазистационарных коррелированных помех — 20–30%;
- некоррелированных шумоподобных помех — 8–10%
- нестационарных помех — 5–7%
- реверберации речи — до 10%.

Синтез речи

В настоящее время наибольшее развитие получили методы синтеза речи на основе артикуляционного [6], лингво-акустического подходов и синтеза по правилам [7]. Основу этих методов составляют достаточно подробная математическая модель артикуляторного тракта, модели речеобразования по разным речевым элементам (фонемам, аллофонам, слогам, дифонам, трифонам и т.д.), а также различные рекуррентные модели с линейным предсказанием. При адекватном наборе исходных элементов артикуляционный и лингво-акустический подходы обеспечивает качественное воспроизведение спектрального состава речи, а набор правил — возможность формирования её естественного просодического оформления.

Проблема состоит в достижении натуральности звучания, приближающейся к естественной речи, устранении недостатков существующих методов при стыковках элементов речи между собой, управлении просодическими характеристиками формируемого сообщения, модификации индивидуальных особенностей синтезируемого голоса. Требуется также углубление лингвистических и акустических знаний о процессах речеобразования для уточнения и расширения набора правил, используемых при управлении процессом синтеза.

Для повышения качества синтеза речи первоочередной задачей является детальное теоретическое и инструментальное исследование формирования динамических процессов речеобразующего тракта, связанных с эффектами коартикуляции, переходом от звука к звуку. Необходимо также определение физических, лингвистических и психологических параметров, создающих натуральность синтезируемой речи. В настоящее время почти не изучены

механизмы восприятия синтетической речи. Поэтому более или менее удачные коммуникативные, модальные, стилевые и эмоциональные интонации в программном синтезе получаются пока не на основе познанных закономерностей, а скорее интуитивно или методом подбора параметров.

Пока предлагаемые решения в большинстве своём служат только стартовыми позициями в решении проблемы, что отражает наше сегодняшнее представление о речеобразовании как системе в целом. Без исследования процессов формирования естественной спонтанной речи практическое применение синтетической речи ограничено.

Литература

1. *Галяшина Е.И.* Слуховая перцепция как базовый метод фоноскопии. Речевые информационные технологии, 2003.
2. *Reynolds D.A., Rose R.C.* Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. // IEEE Trans. on Speech and Audio Proc., 1995. Vol.3, №.1. pp. 72–83.
3. *Потапова Р.Г.* Речь: коммуникация, информация, кибернетика. М.: Радио и связь, 1997. 528с.
4. *Уидроу Б., Стирнз С.* Адаптивная обработка сигналов. М.: Радио и связь, 1989. 440с.
5. *Маркел Дж.Д., Грей А.Х.* Линейное предсказание речи. М.: Связь, 1980. 308с.
6. *Сорокин В.Н.* Синтез речи. М.: Наука, 1992.
7. *Darrow B.* Research Spurs Development of Talking Machines // Design News, 1984. V. 40, №.12.

Максимов Е.М. —

1952 г.р., доктор технических наук. Государственное учреждение «Войсковая часть 35533».

Ромашкин Ю.Н. —

1953 г.р., кандидат технических наук. Московский государственный институт радиотехники, электроники и автоматики (технический университет).

Лопатина С.А. —

1961 г.р., кандидат физико-математических наук. Государственное учреждение «Войсковая часть 35533».

Современные речевые технологии — новое поколение

Итоги семинара

А.С. Нариньяни

19 июня в МЭИ прошёл семинар «СОВРЕМЕННЫЕ РЕЧЕВЫЕ ТЕХНОЛОГИИ (РТ) — НОВОЕ ПОКОЛЕНИЕ». В семинаре приняли участие более тридцати человек, представляющих более 20 компаний и научных групп. В его программу было включено 14 докладов научных и коммерческих коллективов из Москвы, Санкт-Петербурга, Томска и Минска.

Наряду с небывало большим участием речевых докладов в программе недавно прошедшей конференции «Диалог-2008» (http://www.dialog-21.ru/dialog2008/) материалы семинара подтвердили важные тенденции в положении дел этого сектора ИКТ:

1. Несмотря на затянувшийся период кризиса, связанного с отсутствием поддержки этой стратегически важной области НИОКР как со стороны государства, так и крупных отечественных инвесторов, большая часть профессиональных групп сохранили коллективы и уровень работ, который в целом можно характеризовать как мировой.

2. В исследования включилось множество академических и вузовских научных коллективов, проекты которых существенно расширяют фронт работ фундаментального характера.

3. Наряду с заметным повышением качества традиционных компонентов анализа и синтеза речевой информации, очевидна быстро растущая роль двух важнейших составляющих:
- формирования речевых корпусов различной специализации и
- исследования всех уровней просодии, а также всё более высокий уровень разработки специализированных аппаратных и микропроцессорных средств.

В целом обзор представленных результатов и последовавшие обсуждения позволяют сделать выводы:

◆ Общий национальный потенциал работ в области РТ позволяет ставить вопрос о комплексном проекте, способном при достаточном уровне финансирования решить ряд ключевых стратегически важных для страны прикладных задач в сроки от трёх до пяти лет.

◆ Ядром такого проекта может стать Консорциум «Российские Речевые Технологии», основной функцией которого должна стать разработка оптимального плана работ,

ориентированных на чётко определённые задачи в реальные сроки, и формирование эффективной оргструктуры, способной обеспечить реализацию этого плана.

- В качестве важнейших составляющих проект должен включать интеграцию усилий участников на формировании системы рабочих стандартов РТ и создании общего национального речевого корпуса русского языка.

- Успех проекта определяется способностью участников Консорциума найти и внедрить в комплекс работ инновационные технологии следующего поколения, обеспечивающие качественно новый уровень получаемых результатов.

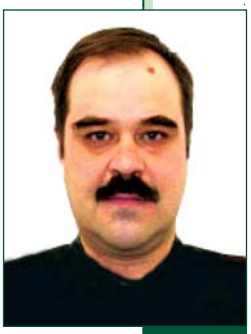
Параллельно с подготовкой семинара шла проработка концепции комплексного проекта, которая будет продолжена в ближайшее время.

Пользуюсь случаем от имени участников семинара выразить благодарность за помощь в его проведении заведующему кафедрой ВТСС МЭИ профессору И.И. Ладыгину и редакционному директору ИТ-группы изданий СК ПРЕСС Э.М. Пройдакову.

Нариньяни Александр Семёнович —

область научных интересов — искусственный интеллект.
Кандидат физико-математических наук, член-корреспондент РАЕН (1997), академик РАЕН (1999).

Результаты работы первого семинара «Обеспечение расследования, раскрытия и профилактики преступлений с использованием фоноскопических экспертиз»



А.Н. Слепич,

И.В. Рыжкова

В соответствии с планом мероприятий МВД России по подготовке и проведению Международного салона средств обеспечения безопасности «Комплексная безопасность-2008», утверждённом 24.11.2007, ГУ НПО «СТиС» МВД России совместно с ЭКЦ МВД России, провели семинар «Обеспечение расследования, раскрытия и профилактики преступлений с использованием фоноскопических экспертиз».

Проведение семинара направлено на поиск перспективных научно-технических решений для проведения фоноскопических исследований и мониторинг разработок в области речевых технологий.

В семинаре приняли участие ведущие организации, проводящие разработки в области речевых технологий, такие как Институт проблем передачи информации им. А.А. Харкевича РАН, Московский государственный лингвистический университет, Московский технический университет связи и информатики, Московский технический университет им. Н.Э. Баумана, ООО «Центр речевых технологий», ЗАО «НПП «ИСТА-Системс», Рязанский государственный радиотехнический радиоуниверситет, ООО «ЮНИКОР микросистемы», ООО «Скинер». Среди участников — 6 докторов и 12 кандидатов наук.

На семинаре были представлены доклады по трём направлениям развития речевых технологий, результаты которых используются в МВД России, это:

— технические средства для расследования преступлений с использованием фоноскопической экспертизы;

— средства контроля доступа на основе использования речевых технологий;

— преобразователи речи.

Расследование, раскрытие и профилактика преступлений основываются на возможности получения доказательной базы, предоставляемой фоноскопической экспертизы, эффективность которой определяется уровнем развития соответствующих технических средств.

На семинаре отмечено, что наиболее перспективным научно-техническим решением для проведения фоноскопической экспертизы на сегодняшний день является подход, использованный сотрудниками ЭКЦ МВД России, заключающийся в создании единого «типового технологического процесса» проведения экспертизы.

Этот подход получил развитие в разработке АПК «Сапфир». «Типовой технологический процесс» представлен в виде последовательности, автоматизированных основных этапов (видов) работ, выполняемых экспертом в процессе производства фоноскопической экспертизы.

Основными этапами работ являются:

— установление дословного содержания исследуемой фонограммы (путём многократного прослушивания фонограммы, дифференциации и атрибуции реплик участников разговора);

— сегментация речевых сигналов интересующих эксперта лиц;

— определение технических параметров и лингвистических характеристик речи исследуемых лиц с целью решения вопросов пригодности речевых сигналов для проведения акустического и лингвистического анализа;

— лингвистический анализ речи исследуемых лиц (определение индивидуальных и групповых признаков речи на уровне речевого потока, уровне фразы и слова и фонетическом уровне). Особую роль при проведении раздельного и сравнительно лингвистического анализа играют признаки фонетического уровня, которые могут даже при минимально допустимой длительности речевых фрагментов составить основную часть индивидуальной совокупности признаков речи исследуемого диктора. Анализ признаков фонетического уровня включает в себя сегментацию сопоставимых речевых фрагментов (как правило, триад звуков), которые используются для сравнения экспертом внутридикторских вариантов произнесения той или иной фонемы и для сопоставления вариантов произнесения данных фонем разными дикторами методами аудитивного и лингвистического анализа. Сопоставимые триады, подвергшиеся лингвистическому анализу, в дальнейшем используются для проведения акустического анализа;

— проведение акустического анализа сигнала: интегрального (макроанализ) и сегментного (микроанализ). На практике при проведении акустического анализа возникают ситуации, где решающую роль для принятия решения о сходстве или различии исследуемых дикторов играют признаки микроанализа звуков — наиболее трудоёмкого вида анализа, сопровождающегося поиском экспертом в спорных фонограммах идентичных речевых фрагментов (триад звуков) и измерении для них спектральных и временных характеристик.

Аппаратурная реализация «типового технологического процесса» проведения фоноскопической экспертизы в АПК «САПФИР» обеспечила следующие функциональные возможности:

1. Производить ввод/вывод сигнала из файла и со звуковой платы или других устройств.

2. Одновременно открывать несколько звуковых файлов (до 60).

3. Получать визуальное отображение сигнала, содержащегося в открытом звуковом файле или введённого с внешнего устройства (сигналы визуализируются в так называемых «Окнах осциллограммы», таких окон может быть открыто

несколько, в одном «Окне осциллограммы» может отображаться несколько звуковых файлов, имеющих разную частоту дискретизации).

4. Прослушивать звуковой сигнал в различных режимах (в режиме псевдостерео, в ускоренном или замедленном темпе без изменения тембра голоса, прослушивать сигнал, преобразованный различными способами).

5. Проводить различные виды спектрального и временного анализа сигнала (вычислять и визуализировать мгновенный спектр, средний спектр, спектрограмму, кепстрограмму, траектории формант, контур основного тона) — для визуализации результатов анализа предусмотрен удобный интерфейс, ориентированный на типичные действия, выполняемые экспертом-фоноскопистом.

6. Проводить исследование фонограммы на предмет наличия или отсутствия следов монтажа (путём исследования непрерывности фазовой составляющей выбранной гармоники спектра, непрерывности спектральных характеристик фоновых шумов, анализ спектрального состава сигнала фонограммы).

7. Создавать описание дословного содержания фонограммы: предусмотрен интерфейс для проведения сегментации фонограммы (на временные интервалы), составления дословного описания содержащихся в сегментах реплик и классификации реплик по дикторам. Введённое через АПК «САПФИР» дословное описание фонограммы сохраняется в файле, имеющим текстовый формат, и может быть прочитано, например, текстовым редактором Microsoft Word с отображением результатов сегментации и атрибуции сегментов по дикторам.

8. Просматривать дословное содержание фонограммы, помещённое в таблицу текстовых интервалов, и/или отображаемое под осциллограммой сигнала синхронно во времени.

9. Автоматически прослушивать и отображать в окне осциллограммы временной интервал, соответствующий выбранной экспертом реплике фонограммы.

10. Проводить автоматический поиск слов или их частей (триад) в дословном содержании фонограммы.

11. Создавать в автоматическом режиме на основе исходной фонограммы и составленного её дословного содержания новый звуковой файл, содержащий реплики выбранного экспертом диктора («дикторский» звуковой файл).

12. Копировать в автоматическом режиме результаты редактирования дословного содержания «дикторского» звукового файла в текстовый файл, содержащий дословное содержание исходной фонограммы.

13. Проводить автоматизированный поиск идентичных фрагментов речи в двух спорных фонограммах (алгоритм поиска ориентирован на то, что исследуемые фрагменты речи будут соответствовать триадам звуков, например, согласный — гласный — согласный).

В АПК «САПФИР» предусматривается система подсказок, помогающая эксперту выбрать оптимальный список параметров описания конкретного речевого фрагмента. АПК рассчитана на экспертов, обладающих самой разной квалификацией. Также следует отметить, что в связи со значительным увеличением в последнее время количества материалов, содержащих речь не на русском языке, предоставляемых для проведения фоноскопической экспертизы, необходимость в проведении автоматизированного поиска идентичных речевых фрагментов для решения задач идентификации иноязычных дикторов может стать особенно острой.

До недавнего времени использование речи в системах средства контроля доступа носило только вспомогательный характер. На прошедшем семинаре доложены результаты использования новых методов анализа, обеспечивающих верификацию диктора с достоверностью выше 99%. Реализация такого верификатора рассчитана на использование в виде «пароля» числительных русского языка. Количество числительных в «пароле» — десять. Для предотвращения несанкционированного доступа «пароль» меняется случайным образом при каждом обращении к системе.

В отдельных применениях для сокращения времени произнесении «пароля» число слов в нём может быть от трёх до семи. Для повышения достоверности число слов может быть увеличено до двенадцати.

Верификатор устойчив к стационарным шумам при SNR выше 12 дБ и не чувствителен к случайным помехам высокой интенсивности, в виде посторонних разговоров и музыки. Предусмотрена защита этого средства контроля доступа от попытки «вторжения» с использованием записи и последующего воспроизведения чужого голоса.

Разработка проведена в ИППИ им. А.А. Харкевича РАН. Испытание речевого средства контроля доступа на базе данных, включающей 429 дикторов (в объёме 125 часов речи), показали, что суммарная ошибка (несанкционированный пропуск и ошибочный отказ в доступе) для 90% дикторов составила менее 0,01%.

Вопросы преобразования речевой информации в текст были представлены на семинаре сотрудниками ООО «СКИНЕР», разработавшими в 2007 году систему автоматизированного преобразования русскоязычной речи в письменный вид.

При разработке систем автоматического распознавания речи и преобразование её в текстовый формат возникает проблема вариативного произношения слов (как одним человеком в различных эмоциональных состояниях, так и разными людьми). Кроме того, на входящий сигнал влияют многочисленные факторы, такие как окружающий шум, отражение, эхо и помехи в канале с заранее неизвестными параметрами.

Система распознавания состоит из следующих взаимосвязанных и взаимодействующих модулей:

— модуля ввода произнесённых слов;
— модуля настройки на особенности речи пользователя, формирования цифровых моделей вводимых слов и ведения словаря;
— модуля распознавания произнесённых слов;
— интеллектуального интерфейса.

Разработанная система распознавания речи основана на фонемно-ориентированном методе.

Разработанная система распознавания речи имеет следующие возможности:

1. Не требуется предварительная настройка системы на пользователя, что существенно упрощает эксплуатацию.

2. Реальная точность распознавания слов при однократном произнесении составляет не менее 75% при объёме словаря до 7000 слов.

3. Не требуется написание строгой грамматики для распознавания речи. Позволяет распознавать не только отдельные слова, но и предложения ограниченного размера.

4. Затрачивает сравнительно небольшое время на обработку звукового сигнала.

Такая система может применяться в call-центрах, для упрощения обслуживания клиентов, а также эту систему можно использовать в службах безопасности организаций.

Мониторинг существующих технических средств для проведения фоноскопической экспертизы показал, что

— положительный опыт использования технических средств, разработанных специально для использования в ЭКП ОВД России;

— появление систем, автоматического поиска записи речи известного лица (диктора) в сколь угодно больших массивах фонетического материала.

Отличительными особенностями технических средств, разработанных для использования в ЭКП ОВД России, является высокий уровень их автоматизации при выполнении идентификационных исследований по голосу и речи; автоматизация при проведении технического исследования фонограмм (идентификация средств звукозаписи, установление аутентичности фонограмм речи); возможность использования различных видов визуализации при проведении анализа звуковых сигналов.

Положительно отмечены АПК «Учёт-Ф», АПК «Фонексия» и АПК «Виртуоз», нашедшие своё применение в ЭКП МВД России.

АПК «Учёт-Ф» позволяет формировать и хранить базы данных, содержащие кодовые образы особенностей устной речи, установочные данные и фонограммы речи объектов, а также позволяет осуществлять поиск в учётном массиве объектов, характеристики речи которых (по установленным критериям) идентичны характеристикам объекта.

АПК «Учёт-Ф» позволяет:

— производить централизованный учёт для нужд МВД, производящих расследование по уголовным делам;

— проводить фоноскопические исследования в интересах оперативно-розыскных подразделений.

АПК «Фонексия» предназначен для автоматизированной диагностики акцентов и диалектов русской устной речи. Этот комплекс используется для проведения фоноскопических экспертиз, направленных на диагностику места рождения или места длительного проживания неизвестного субъекта (диктора), фонограмма речи которого исследуется.

АПК «Виртуоз» предназначен для поиска на речевых фонограммах признаков нарушения их аутентичности, в частности, следов монтажа фонограмм, а также следов цифровой обработки фонограмм и следов выборочной звукозаписи.

Системы автоматического поиска записи речи известного лица (диктора) в сколь угодно больших массивах фонетического материала могут использоваться для поиска лиц, представляющих оперативный интерес и для систематизации постоянно накапливающегося фонетического материала.

К таким системам относятся: «Трал X», имеющая высокую, более чем 100 фонограмм в минуту, скорость поиска, а также «Голос», работающая в реальном масштабы времени.

Дополнительно отмечено, что для аппаратной поддержки АПК проведения фоноскопической экспертизы хорошо зарекомендовали себя устройства для измерения характеристик и формирования электрических сигналов в звуковом диапазоне частот STC-H246 «Камертон». Устройство «Камертон» сертифицировано как средство измерения.

Выводы по результатам работы семинара

1. По известным техническим средствам проведения фоноскопических экспертиз положительную оценку заслуживают: АПК «Виртуоз», АПК «Учёт-Ф», АПК «Фонексия»; дополнительно могут быть рекомендованы к внедрению в службы ЭКП МВД России программное средство «Ривьера» и АПК «Сапфир».

2. Расширение возможности, снижение времени проведения фоноскопической экспертизы связаны с разработкой технических средств, основанных на «технологии проведения фоноскопической экспертизы», объединяющей в единое целое все этапы экспертизы». Первой разработкой, намеченной к внедрению в ЭКП ОВД в 2008 году, является АПК «Сапфир».

3. Безусловным достижением речевых технологий является впервые обеспеченная возможность использования голоса человека для верификации в системах ограничения доступа. Достигнутый уровень достоверности превышает 0,99.

4. Отмечена перспективность начатых разработок в области создания преобразователей русской речи в текст, хотя используемые в настоящее преобразователи имеют недостаточные функциональные возможности, ограничивающие их практическое применение.

5. Перспективными направлениями речевых технологий, с точки зрения использования подразделениями ОВД России, являются:
— создание автоматических средств проведения фоноскопической экспертизы, направленной для расследования преступлений;
— создание средств и технологий проведения лингвистической экспертизы;
— разработка преобразователей речи на основе нейроинформационных технологий.

В целях постоянного отслеживания достижений в области речевых технологий, оценки возможности использования их результатов в МВД России, а также апробации результатов ведомственной науки принято решение о придании настоящему семинару статуса постоянно действующего.

Слепич Андрей Николаевич —

начальник отдела научно-исследовательского института специальной техники ГУ НПО «СТиС» МВД России, образование высшее.
Контактный тел. (495) 673-91-49. e-mail: an@slepich.ru

Рыжкова Ирина Владимировна —

старший научный сотрудник Научно-исследовательского института специальной техники ГУ НПО «СТиС» МВД России, Образование высшее.
Контактный тел. (495) 673-91-06. e-mail: irina-ryzhkova@inbox.ru

Речевые технологии на CeBIT 2008

М.В. Хитров,
генеральный директор компании
«Центр речевых технологий»

Анализ компаний, представленных на выставке CeBIT 2008, занимающихся разработкой и развитием речевых технологий, позволил выявить приоритетные направления этого рынка. Появляется всё больший спрос на системы голосовой биометрии. Этот интерес исходит прежде всего от интеграторов, которые уже сейчас представили на рынке системы верификации по отпечатку пальца и радужной оболочке глаза. Ниша голосовой биометрии относительно свободна и рынок только начинает формироваться. Те системы, которые существуют у конкурентов, явно проигрывают по качеству и надёжности системе Voice Key, разработанной компанией ЦРТ. Сотрудники Aixvox, посетившие стенд ЦРТ, отметили, что Voice Key отличается хорошей шумоустойчивостью, т.е. сохраняет заявленную надёжность даже в условиях повышенного шумового фона, и скоростью работы.

Ещё один быстроразвивающийся рынок речевых технологий — системы синтеза речи и системы распознания команд на основе пофонемного распознавания. Сегодня существуют системы хорошего качества TTS (text-to-speech) и STT (speech-to-text) для основных европейских языков — английского, немецкого, французского, испанского, португальского: для этих языков рынок уже сформировался, и лидирующие позиции заняли компании, имеющие готовые решения для разработчиков. Лидирующий игрок — Nuance, постепенно упрочивает свои позиции компания Loquendo. На выставке CeBIT компания Nuance выставлялась в павильоне навигации и решений для автомобилей: у них был маленький стенд и отсутствие всякого демонстрационного материала. Это свидетельствует о том, что Nuance главным образом работает на рынке B2B, т.е. на рынке конечных потребителей, она практически не работает. В павильоне было много компаний, которые демонстрировали голосовое управление для GPS навигаторов и автомобильных компьютеров, вплоть до своего рода «мобильных офисов» (InPhoDrive. Talking information) http://www.inphodrive.com/home.html.

Сотрудники ЦРТ протестировали систему распознания команд и синтеза речи у InPhoDrive — надёжность очень высокая (около 99%). При этом сама компания только появилась и выходит на рынок (10 человек, 1 год). Также на рынке речевых технологий для автопрома есть компании-интегратор, например, Voice Insight (http://www.voice-insight.com). Они стали интеграторами для компании, которая выпускает системы управления проигрывателем и навигатором для автомобиля. При этом их работа — это выбор движка (Nuance или Loquendo), встраивание в интерфейс. Представитель бельгийского подразделения отмечает, что компания готова работать с русским языком. По его словам, компания Nuance имеет движок для русского языка, но они его не демонстрировали.

Несмотря на ведущие позиции компаний, которые уже имеют TTS и STT, многие интеграторы ищут альтернативные решения. Компания LinguaTEC встраивает движок ViaVoice компании IBM в свои продукты — распознание слитной речи (рынок медицинских и юридических услуг), синтез речи, переводчики. Синтез, который смотрел сотрудник ЦРТ для немецкого и английского языков, отмечен им очень высоко. Для распознания слитной речи, по словам представителей компании, требуется около одного часа тренировки на диктора. Но немецкую речь нашего сотрудника программа не распознала.

В целом сложилось впечатление, что рынок для синтеза качества естественной речи, пофонемного распознания голосовых команд и распознавание слитной речи уже сложился, многие компании разрабатывают и предлагают готовые решения. Ниша русского языка (а также других славянских языков) свободна, и будут успешны те компании, которые в скором времени предложат готовые решения. Интерес вызван большой диаспорой русскоязычного населения в Европе, Израиле и наличие хорошего рынка в России.

В области дикторонезависимого распознавания русского языка Центр речевых технологий обладает весьма продвинутыми разработками, что даёт возможность оформлять технологии распознавания в форме рыночного продукта (готового для встраивания в продукты для конечного пользователя).

Спрос на голосовое управление компьютером отмечается со стороны компаний и организаций, разрабатывающих и продающих продукты для людей с ограниченными возможностями.

Общая картина на рынке технологий распознавания характеризуется небольшим количеством компаний, имеющих собственные разработки алгоритмов распознавания. При этом такие компании практически не занимаются разработкой решений для конечного пользователя. Разработчики технологий распознавания речи поставляют на рынок компоненты, готовые для встраивания в разнообразные системы и программные продукты сторонних производителей. Наблюдаются ситуации, когда между компанией-разработчиком технологии распознавания и компанией-разработчиком готового решения существует компания-интегратор, задача которой — подбор технологий распознавания, удовлетворяющих требованиям продукта для конечного пользователя. Например, для встраивания системы голосового управления в автомобильные электронные устройства компания-производитель таких устройств пользуется услугами интегратора, который берёт на себя исследование технологий распознавания и встраивание их в продукт для конечного пользователя.

Отметим, что на современном рынке технологий распознавания почти отсутствует интерес к дикторозависимому распознаванию одиночных команд. Большинство демонстрируемых решений не предполагают обучения системы на конкретного диктора. Эта общая тенденция при очевидном повышении комфорта работы с системой накладывает и определённые ограничения — распознавание возможно только для определённого языка.

Из истории исследований преобразования речи

Часть 2

В.Г. Михайлов,
доктор филологических наук

IV этап (после 1987 года)

Этот этап характеризуется интенсивным развитием новых информационных технологий, включая вокодеры в сети Интернет (IP-телефония), ведомственной телефонной связи с защитой информации ТЕТРА и мобильной телефонной связи — МТС. Обзор применённых методов и соответствующих технологий дан в работах О.И. Шелухина, Н.Ф. Лукьянцева (2000) и В.Г. Михайлова (2002). Приведём описание некоторых вокодерных алгоритмов.

IP-телефония

История развития IP-телефонии насчитывает около пяти лет. В 1994 г. во время полёта космического челнока Endever агентство США NASA передало по сети Internet изображение корабля и голос космонавта. В следующем году несколько зарубежных фирм предложили кодеки для речевой связи в Internet (один из первых — Internet Phone фирмы VocalTec, Израиль). Наконец, в 1996г. фирма Dialogic (США) совместно с упомянутой фирмой разработалf шлюз — устройство сопряжения абонентов сети общего пользования с коммутацией каналов ТфОП с пользователями сети с коммутацией пакетов Internet и тем самым положили начало голосовой связи в последней — IP-телефонии[1] [Михайлов 2002]. Вскоре появились шлюзы фирм Vienna Systems, Inter-Tel, Cisco и др. Первые шлюзы, которые использовались в России, были изготовлены фирмой Dialogic, а установку шлюзов выполняли фирмы Tario, RGC, Comptek и др. Фирма Tario к 1999 г. установила оборудование в 30 городах России и СНГ и через шлюз в США обеспечила выход пользователей VoIP на все страны мира. Фирма RGC имеет собственные серверы в Москве, Санкт-Петербурге, Владивостоке, Хабаровске, Новосибирске, Южно-Сахалинске, Нью-Йорке, Мюнхене, Берлине.

[1] **IP-**Internet Protocol; **VoIP —** IP-телефония; **ISTP —** Internet Short Time Providers (провайдеры сети IP-телефонии с малым временем замедления); **ITSP —** Internet Telephony Service Providers (провайдеры сети IP-телефонии с предоставлением телефонных услуг).

Сеть **IP**-телефонии обслуживает передачу мультимедийных приложений. К ним относятся речь, ФМ, передача данных и видео.

Оборудование сети IP-телефонии обеспечивает:

1. Кодирование, компрессию и упаковку речевого сигнала в IP- пакеты;

2. Управление потоками IP-пакетов в сети Internet;

3. Интерфейс с телефонными сетями общего пользования ТфОП и сетями подвижной связи, включая сотовые сети.

Эти функции реализованы в виде плат и программного обеспечения для персонального компьютера пользователя (звуковая карта, кодек) и оборудования провайдера (шлюзы, маршрутизаторы и пр.). В 1998 г. Госкомсвязь России официально отнёс услуги IP-телефонии к телематической службе пакетных голосовых соединений. Уже выдано около 700 лицензий отечественным и зарубежным фирмам на развитие и внедрение этой службы (фирмы **CISCO Systems, RGC** (США), **VocalTEC** (Израиль) и др.). В 1999 г. были провайдеры в 30 городах России и СНГ, которые обеспечивали выход по сети IP-телефонии на многие страны мира. К их числу относятся провайдеры сети Tario.Net (оборудование фирмы DM3), объединяющей 55 городов России и стран СНГ и имеющей выход на 237 стран мира. ЗАО «Корпорация ОСС» имеет собственные серверы в Москве, С-Петербурге, Нью-Йорке и более 50 узлов на территории СНГ, обеспечивает соединения через сети ITXC, Teleglobe по всему миру. Фирма RGC развивает глобальную сеть на базе оборудования фирм Cisco, VGW, имеет серверы в России, США, Германии. В ряде сетей применяется также оборудование шведской фирмы Ericsson (например,Sitek), фирмы Clarent (сеть «Элвис-Телеком») и других производителей.

Отметим следующие виды услуг IP-телефонии:

1. Дистанционное обучение по схеме: офисный мультимедийный компьютер — персональный компьютер дистанционно удалённого пользователя или обычный телефон.

2. Деловые местные, междугородние или международные переговоры, включая конференц-связь из четырёх, восьми и более абонентов с обычных телефонов по тарифам, значительно более низким по сравнению с обычной телефонной связью.

3. Продажа, консультации и коммерческая реклама в сети IP-телефонии и в сети ТфОП.

4. Связь территориально разнесённых офисов фирмы для пересылки финансовых отчётов, собраний, деловых обсуждений.

5. Голосовой сервис, в том числе голосовая почта (запись и воспроизведение входящих речевых сообщений), чтение текстов, распознавание устной речи, режим «белая доска» и т.д.

Исходя из потенциальных возможностей технологии IP-телефонии, следует ожидать в последующие годы быстрого развития и совершенствования её средств с постепенным замещением и вытеснением услуг традиционной телефонии.

По оценкам зарубежных специалистов (фирма **Philips Group**) объём услуг рынка IP-телефонии в ближайшие годы вырастет в мире от одного миллиарда до 20–30 миллиардов долларов США в год.

Вместе с тем ещё не решены до конца вопросы качества речевой связи, организации конференц-связи, совместимости технологического оборудования и программного обеспечения разных фирм… Ожидается много трудностей на пути развития IP-телефонии в России: значительная стоимость офисного оборудования и программного обеспечения (порядка 30 тысяч долларов США на 10 абонентских оконечных пунктов), ограниченная пропускная способность многих из действующих каналов связи, несовместимость установленного и нового оборудования. При этом уровень цифровизации основных междугородних и особенно местных телефонных линий в России остаётся кране низким. (По данным компании «Связьинвест», соответствующие показатели в 2000г. составляли 70% и 28,5%. Отметим, что в России ещё в 2003 г. почти 54 тыс. населённых пунктов вообще не были телефонизированы.)

Аппаратные и программные средства VoIP

В качестве абонентских оконечных пунктов сети VoIP выступают ПК, оборудованные звуковой картой типа AWE-64 (Advance Wave Effects) фирмы Creative Labs (США), кодеком (картой или ПО) и модемом со скоростью передачи не менее 14.4 кб/с. ПК подключается через IP-провайдера к сети VoIP. Абоненты сети ТфОП включаются в сеть VoIP также через IP-провайдеров, шлюзы (Gateway) которых обеспечивают оцифровку, кодирование и упаковку речевого сигнала в IP-пакеты. В небольших и корпоративных сетях VoIP маршрутизацию и контроль соединений обеспечивают серверы под управлением стандартного ПО (например, Unix, Windows NT). В крупных сетях для этой цели используют специальные сетевые средства — маршрутизаторы (Gatekeeper), которые обеспечивают необходимую производительность, качество и надёжность соединения.

Взаимодействие мультимедийных оконечных устройств определено рекомендацией H.323 Международного союза электросвязи ITU (International Telecommunication Union). Выработаны соответствующие рекомендации для кодеков речевого сигнала на скорости передачи 6,3; 8,0; 13,0 и 32,0 кб/с (протоколы G.723, G.729, G.728, G.726). Следует отметить конкурирующие предложения разработчиков стандартов для сети Internet по использованию протокола SIP (Session Initiation Protocol — протокол по управлению вызовами и коммутацией), который позволит централизовать все голосовые службы сетей с пакетной коммутацией и положит конец традиционной телефонной системе. По запросу пользователя IP-сети появится возможность поиска абонента в сети, упростятся решения вопросов конференц-связи и совместимости оборудования разных производителей. Протокол SIP способен обслуживать любые виды приложений и коммуникации в реальном времени поверх протокола IP.

Качество речевой связи. Кодеки

К достоинствам сети VoIP следует отнести низкую по сравнению с сетью ТфОП стоимость междугородных и международных соединений, поддержку мультимедийного трафика и развитого сервиса. Вместе с тем следует отметить и присущие сети VoIP недостатки, важнейший из которых связан с качеством передачи. Рассмотрим этот вопрос подробнее.

В сети с коммутацией сообщений время прохождения пакетов оказывается непостоянным: пакеты на ЦКС выстраиваются в очередь, время прохождения зависит от нагрузки в сети ПД; замедление пакетов в виртуальном канале переменно и определяется их маршрутом в сети. В результате время прохождения пакетов по сети **VoIP** может составлять значительную величину (одна и более секунд при использовании каналов спутниковой связи) и сопровождаться нарушением порядка и интервала между временем получения пакетов (т.н. джиттером). Кроме того, некоторая часть пакетов оказывается потерянной, так как для уменьшения времени передачи голосовых данных УЗО не используют. Условно считается допустимым потеря до 5% пакетов, при этом вероятность потери подряд 2–3 пакетов достаточно мала. При задержке, превышающей 150 мс, требуется принимать меры против возникновения в разговорном тракте явления эхо (абонентская сеть ТфОП в основном состоит из двухпроводных линий). Эхозаградители, а также детекторы речь/пауза, применяемые в сетях с уплотнением, вносят свою долю в искажение речевого сигнала.

Качество предоставленных услуг сервиса обслуживания **QoS** (Quality of Service) в цепи «точка — точка» системы пакетной коммутации мультимедийных приложений **ATM** (Asynchronous Transform Mode) определяется долей потерянных пакетов, временем задержки и джиттером. В соответствии с рекомендацией **G.114** ITU качество предоставленных услуг **QoS** классифицируется так.

0…25 мс — малая задержка, качество предоставленных услуг **QoS** по оценке методом мнений **MOS (**Mean Opinion Scores) равно пяти ;

25…150 мс — нормальное интерактивное взаимодействие между пользователями. **MOS** = 4…5;

150…400 мс — эффективное взаимодействие затруднено, но ещё допустимо, требуется эхокомпенсация. **MOS** = 2,9…3,8;

более 400 мс — интерактивное взаимодействие крайне затруднено, требуется эхокомпенсация. Режим полудуплексных переговоров. **MOS** = 2…2,9.

В целом для уменьшения времени задержки голосовых пакетов в сети VoIP современными сетевыми технологиями этим пакетам присваивается приоритет, оптимизируется маршрут и предусматривается определённый резерв по загрузке сети. Джиттер пакетов снижается на порядок при увеличении скорости передачи данных в сети Internet от 28кбит/с до 112 кбит/с.

Для передачи голосового трафика по сети ПД аналоговый речевой сигнал оцифровывается методом ИКМ со скоростью передачи 64 кб/с и кодируется — компрессируется кодеком с помощью методов устранения избыточности. В результате удаётся понизить скорость передачи до 13,0…4,8 кб/с. Отметим, что согласно требованиям к качеству речи в цифровых средствах теле/радиовещания и к хранению аудиостереосигнала на CD для кодирования методом ИКМ используется скорость передачи 1,5 мбит/с, а после преобразования по методу **MPEG** (Motion Picture Experts Group), рекомендованному экспертной комиссией ITU, — до 8 кб/с.

Эффективность использования кодеков в сети **VoIP** растёт при понижении скорости передачи голосового сообщения от 13 к 4,8 кб/с, хотя одновременно увеличивается задержка сигнала в кодеке от 0,1мс до 40…100мс. Поэтому предпочтительны те кодеки, которые обеспечивают минимальную задержку при данной скорости передачи и заданном качестве речи.

Таблица 1

Данные для некоторых видов кодеков

Кодек[2]	Стандарт	Быстрод. DSP, Mips	Скорость, Кб/с	MOS исх	MOS 2 транз.нч
ADPCM	G.72	8	32	3,7	3,49
RPE-LTP (GSM)	ETSI		13	3,5	3,06
VSELP DAMPS	TIA	20	8	3,4	2,53
MP-MLQ	G.72	17	6,4	3,9	3,41

Из приведённых данных наилучшими показателями обладает кодек, реализующий алгоритм многоимпульсной максимально подобной квантизации **MP-MLQ** (Multipulse Maximum Likelihood Quantization). Алгоритм разработан фирмами **AudioCodes** (Израиль) и **DSP Group** (США) для передачи речи со скоростью 4.8, 6.4, 7.2 и 8 кбит/с. В основу алгоритма положен липредер **LPC-10**. При низкой скорости используются алгебраические коды линейного предсказания **ACELP**, при более высокой — **MP-MLQ**. Время задержки — 37,5 мс. В структуре алгоритма поддерживается программирование с плавающей точкой и кодирование с переменной скоростью. Алгоритм даёт возможность снизить скорость передачи ниже 4 кбит/с и получить время задержки 20 мс. Требующаяся скорость обработки составляет 17 mips (млн. инструкций в сек.). Коммерческая реализация алгоритма **MP-MLQ** осуществлена фирмой **RAD Data Communications** (США). Речевой мультиплексер Kilomux — 2000 содержит несколько плат низкоскоростных кодеков KVC –3, обеспечивающих ведение по линии 64 кб/с одновременно 13 разговоров. Использование такого оборудования особенно эффективно на дорогостоящих каналах, например линиях спутниковой связи.

В отличие от других кодеков алгоритм **MP-MLQ** обеспечивает минимальные искажения речевого сигнала при тандемном соединении вокодеров по низкой частоте, например на стыке сети ПД с сетью ТфОП: при двух транзитах по низкой частоте оценка MOS для алгоритма MP-MLQ практически равна оценке для ADPCM (3.41 против 3.49).

Рассмотрим характеристики некоторых кодеков подробнее.

[2] ADPCM (Adaptive Pulse Code Modulation) — адаптивная ИКМ; RPE-LTP (Regular Pulse Excitation-Long Term Prediction — система импульсного возбуждения с долговременным предсказанием; GSM (Global System for Mobile) — европейская система подвижной сотовой связи; VSELP (Vector Sum Exited Linear Prediction) — линейное предсказание с возбуждением векторной суммой, DAMPS (Digital Advanced Mobile Phone) — система сотовой и спутниковой связи США.

Кодеки системы GSM

В системе **GSM** для кодирования речевого сигнала используется регулярное импульсное возбуждение и долгосрочное предсказание **RPE-LTR**. В блоке предварительной обработки осуществляется коррекция (предыскажение) спектра входного сигнала при помощи цифрового фильтра, подчёркивающего верхние частоты. Далее на сегментах по 20 мс производится измерение восьми коэффициентов линейного предсказания *ki*, которые перед передачей в канал связи преобразуются в логарифм отношения площадей *ri*, причём для функции логарифма используется кусочно-линейная аппроксимация. Сигнал с выхода блока предварительной обработки поступает на фильтр — анализатор кратковременного линейного предсказания и по его выходному сигнал-остатку предсказания *eh* оцениваются параметры долговременного предсказания: коэффициент предсказания *g* и задержка *t*. При этом сегмент-остаток из 160 отсчётов кратковременного предсказания *e** разделяется на четыре подсегмента длительностью по 5 мс из сорока выборок в каждом. Параметры *t* оцениваются для каждого из подсегментов в отдельности, причём для оценки задержки *t* для текущего подсегмента используется скользящий подсегмент из 40 выборок, перемещающийся в пределах предшествующих 128 выборок сигнал — остатка предсказания *e**. Сигнал *e** фильтруется фильтром-анализатором долговременного линейного предсказания, а выходной сигнал последнего *fh* фильтруется сглаживающим фильтром, и по нему формируются параметры сигнала возбуждения в отдельности для каждого из подсегментов по 40 выборок. Сигнал возбуждения одного подсегмента состоит из 13 импульсов, следующих через равные промежутки времени (втрое большие, чем интервал дискретизации исходного сигнала) и имеющих различные амплитуды.

Таблица 2

Число бит для кодирования параметров по системе GSM

Передаваемые параметры	Число бит	Примечание
Параметры фильтра кратковременного предсказания (логарифм отношения площадей r_i, i = 1…8)	36	R_1, r_2 — по 6 бит; r_3, r_4 — по 5 бит; r_5, r_6 — по 4 бита; r_7, r_8 — по 3 бита;
Параметры фильтра долговременного предсказания (коэффициент предсказания *g*, и задержка *t* для каждого из четырёх подсегментов)	36	*g* — 2 бита; *t* — 7 бит;
Параметры сигнала возбуждения (номер последовательности n, максимальная амплитуда *v*, нормирование амплитуды импульсов b_i, i = 1…13, для каждого из четырёх подсегментов)	188	*n* — 2 бита; *v* — 6 бит; b_i — 3 бита
Всего	260	

Для формирования сигнала возбуждения 40 импульсов подсегмента сглаженного остатка f_n обрабатываются следующим образом. Последний (сороковой) импульс отбрасывается, а первые 39 импульсов разбиваются на три последовательности по 13 импульсов: в первой — импульсы 1,4,7,…,37, во второй — импульсы 2,5,8,…,38, в третьей — импульсы 3,6,9,…,39. В качестве сигнала возбуждения выбирается та из последовательностей, энергия которой больше.

Амплитуды импульсов нормируются по отношению к импульсу с наибольшей амплитудой. Нормированные амплитуды кодируются тремя битами каждая по линейной шкале квантования. Абсолютное значение наименьшей амплитуды кодируется шестью битами в логарифмическом масштабе. Положения начального импульса 13-элементной последовательности кодируется двумя битами, т.е. формируется номер последовательности, выбранной в качестве сигнала возбуждения для данного подсегмента. Таким образом, выходная информация кодека для 20 мс сегмента речи включает в себя: параметры фильтра кратковременного линейного предсказания — восемь коэффициентов логарифма отношения площадей r_i, i=1,…,8; параметры фильтра долговременного линейного предсказания — коэффициент предсказания q и задержку t для каждого из четырёх подсегментов; параметры сигнала возбуждения, номер последовательности, максимальную амплитуду v, нормированные амплитуды b_i,, i = 1…13, импульсов последовательности для каждого из четырёх подсегментов. Всего для одного 20 мс сегмента речи отводится 260 бит, т.е. кодер осуществляет сжатие информации в пять раз (1280 : 260 = 4,92).

Перед выдачей в канал связи выходная информация кодера подвергается дополнительно канальному кодированию. В декодере блок формирования сигнала возбуждения, используя принятые параметры возбуждения, восстанавливает 13-импульсную последовательность сигнала возбуждения для каждого из подсегментов, включая амплитуды импульсов и их расположение во времени. Сформированный таким образом сигнал возбуждения фильтруется фильтром-синтезатором долговременного предсказания, на выходе которого получается восстановленный остаток предсказания фильтра-анализатора кратковременного предсказания. Последний фильтруется решётчатым (лестничным) фильтром-синтезатором кратковременного предсказания, причём параметры фильтра предварительно преобразуются из логарифма отношений площадей r_i в коэффициенты частной корреляции k_i.

Выходной сигнал фильтра — синтезатора кратковременного предсказания фильтруется (в блоке постфильтрации) цифровым фильтром, восстанавливающим амплитудные соотношения частотных составляющих сигнала речи, т.е. компенсирующим коррекцию, внесённую входным фильтром блока предварительной обработки кодера. Сигнал на выходе постфильтра является восстановленным цифровым сигналом речи. GSM-кодек выдаёт информацию со скоростью 13 кбит/с. Главным для этого кодека является то, что он может быть легко реализован для работы в реальном времени при малых вычислительных затратах.

При передачи в системе **GSM** используется техника прерывистой передачи **DTX** (Dicontinuos Transmission). При такой системе передачи в групповом радиоканале каждый речевой канал активен не непрерывно. В дуплексном режиме переговоров каждый участник говорит менее 50 % времени. Кроме того, во время разговора между словами и фразами также есть паузы. Для определения интервалов активности используется детектор активности речи (**VAD**); при этом групповой канал связи в обнаруженных паузах может быть предоставлен для передачи других переговоров. При использовании **VAD** в паузах речи могут передаваться неречевые данные. Эффективность таких систем зависит от алгоритма **VAD**, который работает в условиях воздействия внешних шумов, типичных, например, для подвижной автомобильной радиосвязи.

Прерывистая передача DTX

DTX — эффективный способ повышения эффективности подвижных систем передачи речи. Основной принцип DTX — включение передатчика только в те временны̀е интервалы, когда присутствует речевой сигнал. Для устранения интерференции с соседними каналами и для

предохранения аккумуляторных батарей носимых радиостанций от разряда передатчик может выключаться. Основной проблемой **DTX** является потенциальное снижение качества речи из-за того, что речь может идентифицироваться как шум. При использовании детектора активности канала возможны следующие нежелательные явления: пропадания участков речи и возможность того, что шум будет неправильно идентифицирован как речевой сигнал. Пропадания могут существенно снизить общее качество речи.

В случае, когда VAD используется для включения и выключения передатчика, шум на приёмной стороне может изменяться по уровню. Это явление связано с тем, что при включённом передатчике фоновый шум передаётся вместе с речью. Однако когда речевой сигнал отсутствует, передатчик выключается, что приводит к снижению фоновых шумов до очень низкого уровня. Это случайное изменение в уровне шумов неприятно для слушателя и может повлиять на разборчивость речи. Для уменьшения этого эффекта на время выключения передатчика в декодере производится генерация шума. Этот шум должен быть похож, например, на шум машины или поезда на передающей стороне. Поэтому передатчик периодически передаёт информацию о среднем уровне фонового шума (т.н. «комфортный шум»).

Передатчик состоит из кодека речи, VAD и измерителя уровня фонового шума. Когда на входе есть речь, передатчик включён. Во время речевых пауз передатчик выключается. Через определённое небольшое время передатчик включается на один фрейм для передачи информации о среднем уровне фона и генерации на приёмной стороне комфортного шума. На приёмной стороне при наличии речевого сигнала происходит нормальный синтез. Если не поступает новой информации о фоне, используются существующие параметры шума и генерируется комфортный шум. Когда на приём поступают новые параметры фонового шума, начинается генерация нового комфортного шума. Обычно на стороне декодера используется индикатор «хороший/плохой» фрейм, чтобы показать, верны или нет декодированные параметры, и если верны, то производится замена фрейма. Эффективность DTX зависит от точности VAD.

Кодеки стандарта DAMPS

Цифровой стандарт мобильной радиосвязи DAMPS (Digital Advanced Mobile Phone Service), принятый в США в 1990 г., по своим функциональным возможностям и предоставляемым услугам приближается к стандарту GSM. Стандарт DAMPS не принят в европейских странах, за исключением России, где он ориентирован в основном на региональное использование. В стандарте DAMPS используется метод кодирования VSELP. Блок предварительной обработки выполняет цифровую фильтрацию входного сигнала с подъёмом верхних частот. Для каждого 20 мс сегмента оцениваются параметры фильтра кратковременного линейного предсказания — 10 коэффициентов частной корреляции r_i, $i = 1…10$, которые непосредственно кодируются для передачи в канал связи без каких-либо дополнительных преобразований, и определяется энергия сегмента речи p. Сигнал с выхода блока предварительной обработки фильтруется фильтром-анализатором кратковременного линейного предсказания $A(z)$, имеющего форму инверсного линейного фильтра,

для чего коэффициенты частной корреляции преобразуются в коэффициенты линейного предсказания a_i.

Выходной сигнал фильтра кратковременного предсказания (сигнал-остаток предсказания e_h) используется для оценки параметров фильтра *P(z)* долговременного предсказания. Оценки даются для каждого из четырёх подсегментов по 40 выборок, на которые разделяется сегмент из 160 выборок. Для каждого из подсегментов определяются параметры сигнала возбуждения. Для этого в составе кодера используется схема, аналогичная входящей в состав декодера, которая включает фильтры-синтезаторы кратковременного *H(z)* и долговременного *R(z)* предсказания и две кодовые книги и реализуется метод «анализа через синтез».

Каждая из кодовых книг сигнала возбуждения содержит 128 кодовых векторов, по 40 элементов в каждом. Все кодовые векторы одной книги являются элементами 7-мерного линейного подпространства в 40-мерном пространстве. Каждая кодовая книга, содержащая 128 векторов, задаётся семью базисными векторами и 128 кодовыми словами (7-элементными векторами коэффициентов линейных комбинаций) с однобитовыми элементами.

Сигнал возбуждения фильтра-синтезатора кратковременного предсказания является суммой векторов возбуждения из двух кодовых книг и вектора с выхода фильтра-синтезатора долговременного предсказания. Векторы возбуждения из кодовых книг до подачи на сумматор умножаются на соответствующие коэффициенты усиления t_1 и t_2, а входным сигналом фильтра — синтезатора долговременного предсказания является, в зависимости от участка сегмента, выходной сигнал того же фильтра или суммарный сигнал возбуждения фильтра-синтезатора кратковременного предсказания. Параметры сигнала возбуждения — номера векторов возбуждения I_1 и I_2 из первой и второй кодовых книг и соответствующие коэффициенты усиления t_1 *и* t_2 — определяются по критерию минимума среднеквадратичной ошибки на выходе фильтра-синтезатора кратковременного предсказания, входящего в состав кодера. Предварительно базисные векторы обеих кодовых книг декоррелируют: для первой книги — по отношению к выходному вектору фильтра-синтезатора долговременного предсказания, для второй книги — по отношению к тому же выходному вектору и к базисным векторам первой книги.

В результате выходная информация кодера речи для 20 мс сегмента включает:

- параметры фильтра кратковременного линейного предсказания — 10 коэффициентов частной корреляции r_i; i = 1…10, и амплитудный множитель p — один выбор на весь сегмент;
- параметры фильтра долговременного линейного предсказания — коэффициент предсказания g и задержку τ — для каждого из четырёх подсегментов;
- параметры сигнала возбуждения — номера I_1 *и* I_2 векторов возбуждения из двух кодовых книг и соответствующие коэффициенты возбуждения t_1 и t_2 для каждого из четырёх подсегментов.

Перед передачей в канал связи выходная информация кодера речи подвергается дополнительному канальному кодированию, причём разные параметры в зависимости от их важности для обеспечения качества речи кодируются с различной степенью избыточности.

Таблица 3

Характеристики кодека (согласно стандарту DAMPS)

Передаваемые параметры	Число бит	Примечание
Параметры кратковременного предсказания (коэффициенты частичной корреляции r_i, i = 1…10)	38	k_1 — 6 бит; k_2, k_3 — по 5 бит; k_4, k_5 — по 4 бита ; k_6,…, k_9 — по 3 бита; k_{10} — 2 бита
Амплитудный множитель (энергия сегмента) p	5	
Задержка фильтра долговременного предсказания t	28	7 бит на каждый подсегмент
Номера векторов возбуждения I_1 и I_2	56	I_1 и I_2 по 7 бит
Коэффициенты усиления g, г$_1$ и г$_2$	32	8 бит на каждый подсегмент; векторное квантование
Всего на 20 мс сегмент	159	

Общий объём информации, выдаваемой в канал связи равен 8000 бит/с. Поскольку исходный объём информации на выходе кодека составляет 1280 бит (160 выборок по 8 бит), кодек осуществляет сжатие информации более чем в восемь раз.

В декодере сигнал возбуждения фильтра-синтезатора кратковременного предсказания формируется таким же образом, как и в синтезирующей схеме кодека: по номерам I_1 и I_2 из кодовых книг выбираются векторы возбуждения, которые умножаются соответственно на коэффициенты усиления t_1 и t_2 и складываются с выходным вектором фильтра-синтезатора долговременного предсказания, определяемого параметрами g и t.

Окончательно сигнал возбуждения фильтруется фильтром-синтезатором кратковременного предсказания, выполненного в виде инверсного фильтра, т.е. параметры фильтра преобразуются из коэффициентов частной корреляции r_i в коэффициенты предсказания a_i. Для улучшения субъективного качества синтезированной речи выходной сигнал фильтра-синтезатора подвергается цифровой адаптивной пост-фильтрации, и с выхода пост-фильтра получается восстановленный цифровой речевой сигнал.

Стандарт TETPA

TETPA разработан как единый общеевропейский цифровой стандарт на основе технических решений и рекомендаций стандарта GSM и ориентирован на создание систем связи, эффективно и экономно поддерживающих совместное использование сети различными группами пользователей с обеспечением засекречивания информации. При разработке стандарта ориентировались, прежде всего на создание профессиональных систем связи, хотя системы связи общего пользования также могут быть созданы на основе этого стандарта.

Системы по стандарту TETRA предназначены для организации связи с абонентами телефонных сетей, радиосвязи, передачи данных. В стандарт входят спецификации беспроводного интерфейса, интерфейсов между сетью TETRA и цифровой

сетью с интеграцией услуг (ISDN), телефонной сетью общего пользования, сетью передачи данных, учрежденческими АТС.

TETRA — полностью цифровая система, поддерживает обслуживание речевых сообщений и данных различного формата с обеспечением выбора скорости передачи данных и уровня защиты от ошибок. В системе используется технология TDMA с четырьмя каналами на одной несущей при разносе между несущими 25 кГц, что обеспечивает высокую эффективность использования частотного спектра.

Время установления вызова в системе составляет 300 мс. TETRA поддерживает работу в режиме полудуплекса для связи групп и дуплекса для индивидуальных вызовов. Возможности группового и циркулярного вызовов соответствуют требованиям большинства пользователей. Схема многих приоритетов обеспечивает эффективное распределение ресурса для самых важных соединений в сети.

Основные характеристики протокола радиоинтерфейса стандарта TETRA. Рабочие частотные каналы отстоят друг от друга на 25 кГц. На каждом частотном канале располагаются четыре временны̀х интервала, которые и являются физическими каналами связи, т.е. основным элементом временной структуры является MVDR-кадр (Minimum Variance Distortionless Responce), который содержит четыре пакета. Скорость передачи данных одного канала (пакета) 7,2 кбит/с. Речевой сигнал кодируется со скоростью передачи 4,8 кбит/с с использованием метода ACELP. Цифровые данные с выхода кодера подвергаются блочному свёрточному кодированию, перемежению и шифрованию, после чего формируются информационные каналы. Пропускная способность одного информационного канала составляет 7,2 кбит/с, а скорость цифрового информационного потока данных — 28,8 кбит/с. При этом общая скорость передачи символов в радиоканале за счёт дополнительной служебной информации и контрольного кадра в мультикадре равна 36 кбит/с.

В стандарте TETRA предусмотрены различные способы защиты пользовательских данных от помех в радиоканале, а также несколько вариантов использования полосы пропускания радиоканала. Если пользователь сам обеспечивает целостность и достоверность пакетов данных, система TETRA может представить «прозрачный» канал связи, не внося дополнительных символов корректирующего кода. В этом случае скорость передачи данных при использовании только одного временно̀го интервала в кадре составит 7,2 кбит/с. Для повышения скорости передачи данных может быть предоставлено от одного временно̀го интервала до четырёх. В этом случае скорость передачи данных пропорционально увеличивается и составляет 28,8 кбит/с. Если пользователь не обеспечивает достоверность информации собственными средствами, можно использовать систему помехоустойчивого кодирования системы TETRA. При этом можно работать с использованием кодовых скоростей 2/3 (4,8 кбит/с — низкая степень защиты) или 1/3 (2,4 кбит/с — высокая степень защиты).

MVDR — кадр содержит четыре временны̀х интервала (пакета). Пакет в кадре соответствует независимому каналу передачи информации. Каждый пакет, в зависимости от его назначения, имеет свою внутреннюю структуру. Пакет содержит 510 бит цифровой информации, что соответствует 255 символам модуляции.

Восемнадцать MVDR-кадров объединены в мультикадры, которые, в свою очередь, образуют гиперкадр длиной 60 мультикадров. В стандарте TETRA для организации связи между подвижным абонентом и базовой радиостанцией предусматривается выделение дуплексной пары радиочастот. Так как при этом используется временно̀е уплотнение до четырёх независимых каналов, для снижения взаимных помех в системе применяется жёсткая синхронизация пакетов подвижных станций относительно пакетов, передаваемых базовой станцией; при этом

последовательность пакетов мобильных станций задерживается на две позиции относительно пакетов базовой станции.

Практически в любом пакете от базовой станции к абонентской имеются поля, предназначенные для передачи команд управления и сигнализации. Таким образом, помимо канала управления в 18-м кадре сигнала базовой станции, элементы управления присутствуют во всех информационных кадрах. Метод линейной модуляции, применяемый в TETRA — t/4 — QDPSK.

Определены классы мощности для радиостанций, используемых в системе: 25, 10, 3 и 1 Вт. Радиостанции могут автоматически регулировать выходную мощность в соответствии с нужной напряжённостью поля. Для увеличения зоны действия носимых или возимых абонентских радиостанций предусматривается их использование как ретрансляторов для выхода во внешние сети для работы в сети или организации локальных сетей в режиме двухчастотного симплекса.

Сети стандарта TETRA предполагают распределённую инфраструктуру управления и коммутации, обеспечивающую быструю передачу вызовов и сохранение локальной работоспособности системы при отказе её отдельных элементов. Основными элементами сетей TETRA являются базовые и мобильные станции, устройства управления базовыми станциями, контроллеры базовых станций, диспетчерские пульты, терминалы технического обслуживания и эксплуатации. Функции сетевого обслуживания и межсистемного взаимодействия определяются специальными интерфейсами:

- радиоинтерфейсом, определяющим взаимодействие базовой станции с мобильными абонентскими радиостанциями;
- радиоинтерфейсом непосредственного соединения между абонентскими радиостанциями;
- интерфейсом проводной связи, связывающим контроллер базовой станции с диспетчерским пультом;
- межсистемным интерфейсом для организации связи между конроллерами базовых станций различных сетей;
- интерфейсом связи между терминалами передачи данных и мобильной станцией или диспетчерским пультом;
- интерфейсом управления сетью;
- интерфейсом для подключения к учрежденческим АТС, ТФОП, ISDN, сети с коммутацией пакетов.

Система стандарта TETRA предусматривает выполнение следующих функций:

- индивидуальный вызов между абонентскими радиостанциями и ведение радиотелефонных переговоров;
- вызов абонентских радиостанций со стороны абонентов телефонной сети и вызов абонентов телефонной сети со стороны абонентских радиостанций;
- групповой вызов;
- аварийный вызов;
- передача данных (коротких сообщений, статусных сообщений и пакетов данных);
- передача циркулярных сообщений;
- организация конференц-связи;
- присвоение приоритетов;
- переадресация вызовов;

- организация динамических групп абонентов;
- защита от несанкционированного доступа к сети;
- постановка на очередь абонентов при отсутствии свободных каналов с последующим обслуживанием;
- установка таймеров длительности вызова, соединения;
- ограничение перечня предоставляемых услуг для отдельных абонентов;
- контроль оборудования.

Данные любого характера и сообщения произвольной длины могут быть переданы в системе TETPA с использованием службы пакетной передачи. Это служба предоставляет пользователям сервис протокола IP. Служба пакетной передачи по протоколу IP позволяет использовать все виды протоколов транспортного уровня. Это может быть как датаграммный протокол (UDP), не предусматривающий установления логического соединения, так и ориентированный на установление сеанса связи. Пользователи могут использовать различное программное обеспечение, работающее с протоколами TCP/IP. При подсоединении к инфраструктуре TETPA внешнего оборудования обработки данных, а также при присоединении к абонентским радиостанциям абонентского терминального оборудования используются протоколы X.25, LAP.B, X.21.

Речевой кодек TETRA основан на модели кодирования CELP — линейное предсказание с кодовым возбуждением. В этой модели блок из N речевых выборок синтезируется путём фильтрации соответствующей обновлённой последовательности из кодовой книги, масштабированной коэффициентом усиления g_c, с помощью двух изменяющихся во времени фильтров. Первый фильтр является фильтром долгосрочного предсказания (фильтром основного тона), цель которого — моделирование псевдопериодического речевого сигнала, а второй — фильтр краткосрочного предсказания — моделирует огибающую речевого спектра.

Передаточная характеристика долгосрочного фильтра (или фильтра синтеза основного тона) определяются формулой:

$$\frac{1}{B(z)} = \frac{1}{1 - g_p \cdot z^{-T}},$$

где T — задержка основного тона; g_p — коэффициент усиления основного тона.

Фильтр синтеза основного тона выполнен как адаптивная кодовая книга, где для задержек, меньших чем длина подфрейма, повторяется последнее возбуждение.

Краткосрочный фильтр синтеза определяется формулой:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + a_i \cdot z^{-1}},$$

где a_i, i=1,..., p, — параметры линейного предсказания; p — порядок предсказателя. В кодеке TETRA p=10.

Для определения основного тона и параметров кодовой книги возбуждения в кодеке TETRA используется способ анализ-через-синтез. При способе анализ-через-синтез синтезированная речь вычисляется для всех кандидатов — последовательностей, составляя особую последовательность, которая и формирует выходной сигнал, наиболее близкий к исходному, в соответствии со взвешенной величиной измеренных искажений. Фильтр взвешивания определяется формулой:

$$W(z) = \frac{A(z)}{\tilde{A}(z)} \quad ,$$

где *A(z)* — обратный (инверсный) фильтр линейного предсказания; (0 < z_1 (используется значение z_1 = 0,85). Для взвешивающего фильтра *W (z)* и фильтра синтеза *H(z)* используются квантованные параметры линейного предсказания.

В алгебраическом CELP (ACELP) используется специальная кодовая книга, имеющая алгебраическую структуру. Эта алгебраическая структура имеет некоторые преимущества в отношении сохранения, сложности поиска и помехоустойчивости (робастности). Кодек TETRA использует специальную динамическую алгебраическую кодовую книгу возбуждения, посредством которой, а также динамической матрицы формы образуются фиксированные векторы возбуждения. Матрица формы — это функция модели *A(z)* линейного предсказания. Главная её роль — формировать векторы возбуждения в частотной области так, чтобы их энергия была сконцентрирована в наиболее важных частотных полосах. Используемая матрица формы является триангулярной Теплицевой матрицей низшего порядка, сформированной из импульсного отклика фильтра:

$$F(z) = \frac{A\left(\frac{z}{z_1}\right)}{A\left(\frac{z}{z_n}\right)} \quad ,$$

где *A(z)* — инверсный фильтр линейного предсказания (в конкретных реализациях z_1 = 0,75 и z_2 = 0,85).

В кодеке TETPA используются фреймы длительностью по 30 мс. Это требуется для того, чтобы параметры краткосрочного предсказания вычислялись и передавались в каждом речевом фрейме. Речевой фрейм разделён на 4 подфрейма по 7,5 мс (60 выборок). Основной тон и параметры алгебраической кодовой книги также передаются в каждом подфрейме. В табл. 4 представлено распределение бит для кодека TETPA. Описание каждого фрейма длительностью 30 мс занимает 137 бит, что в результате даёт скорость передачи параметров речи 4567 бит/с.

Таблица 4

Распределение бит для кодека TETPA

Параметр	Номер сегмента				Всего в кадре
Коэффициенты линейного предсказания	1	2	3	4	26
Период основного тона	8	5	5	5	23
Индекс алгебраической кодовой книги	16	16	16	16	64
Коэффициенты усиления	6	6	6	6	24
Всего					137

Литература

1. *Акинфиев Н.Н.* К вопросу построения речевых сообщений // Доклады комиссии по акустике АН СССР. Апрель 1956.

2. *Архипова А.Д., Сапожков М.А.* Перспективы повышения качества вокодерной речи // Материалы шестой Всесоюз. Акуст. конф. М., 1968.

3. *Вокодерная* телефония / Под ред. А.А. Пирогова. М., 1974.

4. *Волошенко Ю.Я., Михайлов В.Г., Морозов Н.А.* К вопросу о регистрации колебаний голосовых связок // Вопросы радиоэлектроники. 1968. Сер. XI. Вып. 7.

5. *ГОСТ Р 50840–95.* Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. М., 1995.

6. *Дремов А.Н.* Решительный шаг к интеграции // Технологии и средства связи. 2001. № 2.

7. *Калачев К.Ф.* В круге третьем. М., 2001.

8. *Калинцев Ю.К.* Разборчивость речи в цифровых вокодерах. М., 1991.

9. *Коротаев Г.А., Михайлов В.Г.* Синтетическое телефонирование // Радиоэлектроника и электронная техника. 1964.

10. *Коротаев Г.А., Михайлов В.Г.* Современное состояние техники параметрического компандирования речи // Зарубежная радиоэлектроника. 1966. № 4.

11. *Котельников В.А.* Теория потенциальной помехоустойчивости. М., 1956.

12. *Кубышкин Ю.И., Халышкин А.С.* Полосный полувокодер для применения на междугородных линиях связи // Труды VII Всесоюз. Акуст. конф. Л., 1971.

13. *Лейтес Р.Д., Соболев В.Н.* Принципы цифрового моделирования вокодеров // Электросвязь. 1966. № 7.

14. *Литвак И.М.* О разработке систем типа вокодер // Доклады комиссии по акустике АН СССР. Апрель 1956.

15. *Мартынов В.С.* Выделитель основного тона // Доклады комиссии по акустике АН СССР. Апрель 1957.

16. *Масленников И.* Будущее реального времени // Доклады комиссии по акустике АН СССР. Апрель 1957.

17. *Материалы* семинара «IP-телефония и дистанционное обучение». М., 2000.

18. *Михайлов В.Г.* Формантное распределение для мужских голосов //Акустический журнал. 1972. Т. 1.

19. *Михайлов В.Г.* Аппаратурные методы измерения качества телефонной передачи // Зарубежная радиоэлектроника. 1973. № 5.

20. *Михайлов В.Г.* Семинар по речевой связи в Стокгольме // Электросвязь. 1975. № 4.

21. *Михайлов В.Г.* Информационные и статистические параметры устной речи. М., 1992.

22. *Михайлов В.Г.* Новые информационные технологии. IP-телефония // Системы и средства связи, телевидения и радиовещания. 2000. № 3.

23. *Михайлов В.Г.* IP-телефония // Акустика речи и прикладная лингвистика. Ежегодник РАО. Вып. 3. М., 2002.

24. *Муравьев В.Е.* Гармоническая система кодирования речи // Труды гос. НИИ Минсвязи. 1959. Вып. 1(15).

25. *Мясников Л.Л.* Объективное распознавание звуков речи // Журнал технической физики. 1943. Т. 13. № 3.

26. *Пирогов А.А.* Гармоническая система сжатия спектров речи // Электросвязь. 1959.

27. *Покровский Н.Б.* Расчёт и измерение разборчивости речи. М., 1962.

28. *Потапова Р.К.* Основные современные способы анализа и синтеза речи. М., 1971.

29. *Потапова Р.К.* Речевое управление роботом: лингвистика и современные автоматизированные системы. М., 1989.

30. *Потапова Р.К.* Речь: коммуникация, информация, кибернетика. М., 1997.

31. *Рабинер Л.Р., Шафер Р.В.* Цифровая обработка речевых сигналов: Пер. с англ. М., 1981.

32. *Радиовещание* и электроакустика / Под ред. Ю.А. Ковалгина. М., 1998.

33. *Распознавание* слуховых образов / Под ред. Н.Г. Загоруйко, Г.Я. Волошина. Новосибирск, 1970.
34. *Рейман Л.Д.* Россия на пути к информационному обществу // Технологии и средства связи. 2001. № 3.
35. *Росляков А.В., Самсонов М.Ю., Шибаева И.В.* IP-телефония. М., 2001.
36. *Сапожков М.А.* О методах компрессии речи // Электросвязь. 1958. № 8.
37. *Сапожков М.А.* Речевой сигнал в кибернетике и связи. М., 1963.
38. *Сапожков М.А., Михайлов В.Г.* Вокодерная связь. М., 1983.
39. *Фант Г.* Анализ и синтез речи: Пер. с англ. Новосибирск, 1970.
40. *Фланаган Дж.* Анализ, синтез и восприятие речи: Пер. с англ. М., 1968.
41. *Шелухин О.И., Лукьянцев Н.Ф.* Цифровая обработка и передача речи. М., 2000.
42. *Цемель Г.И.* Системы сокращения спектра речевого сигнала // Электросвязь. 1957. № 5.
43. *Atal B.* High — quality speech at low bit rates: multi — pulse and stochastically exited linear predictive coders // ICASSP-86. Tokyo. 1986.
44. *Dolansky L., Tjernlung P.* On certain irregularities of voiced-speech waveforms // IEEE Trans. Audio and El. 1968. V. 16. № 1.
45. *Dudley H.* A synthetic speaker // J. Franklin Inst.. 1939.
46. *Dudley H.* Remaking speech // JASA. 1939. № 11.
47. *Dudley H.* Vocoders // Bell Labs Record. 1939. № 17.
48. *Gold B.* Computer program for pitch extraction // J. Acoust. Soc. Am. 1962. V. 32. № 7.
49. *Halsey R., Swaffield J.* Analysis — synthesis telephony with special reference to the vocoder // J. of the Inst. of El. Eng. 1948. V. 95. № 34.
50. *Koenig W., Dunn H., Locy L.* The Sound spectrograph // JASA. 1946. № 18.
51. *Potter R.K.* Visible speech. N.Y., 1947.
52. *Schroeder M.R.* Vocoders: analysis and synthesis of speech // Proc. of IEEE. 1966. V. 54. № 5.
53. *Schroeder M., David E.* A vocoder for transmitting 10 kc/s Speech over a 3,5 kc/s channel // Acoustica. 1960. V. 10. № 1.
54. *Shannon C.* A mathematical theory of communication // Bell system Techn. J. 1948. V. 27. № 3; № 4.
55. *Specom — 1999.* Proc. of Int. Workshop «Speech and Computer». Moscow, 1999.
56. *Specom — 2001.* Proc. of Int. Workshop «Speech and Computer». Moscow, 2001.
57. *Speech* synthesis / Ed. by J.L. Flanagan, L.R. Rabiner. N.Y., 1973.
58. *Tremain T.* The government standard linear predictive coding algorithms LPC — 10// Speech technology. 1982. V. 1. № 2.
59. *Voice* Compression Technology. RAD data Communications. White paper. Ver. 2. 5/96. Cat. 801105.

Михайлов Вадим Георгиевич —

старший научный сотрудник, доктор филологических наук, Московский государственный университет им. Ломоносова, филологический факультет;
Окончил в 1955 г. Ленинградский институт киноинженеров по специальности «звукозапись». В течение 30 лет работал в научно-исследовательских институтах и с 1985 г. — в Московском государственном университете им. М.В. Ломоносова в должности старшего научного сотрудника; В Московском государственном лингвистическом университете на должности профессора.
Область научных интересов: акустико-перцептивные свойства речевого сигнала, автоматическое распознавание и разборчивость речи. Имеет около 100 научных публикаций, 3 монографии, 12 изобретений, в том числе гос. стандарт по оценке качества передачи речевого сигнала.