**УДК 681.5**

**О.С. Агашин, О.Н. Корелин**

# МЕТОДЫ ЦИФРОВОЙ ОБРАБОТКИ РЕЧЕВОГО СИГНАЛА В ЗАДАЧЕ РАСПОЗНАВАНИЯ ИЗОЛИРОВАННЫХ СЛОВ С ПРИМЕНЕНИЕМ СИГНАЛЬНЫХ ПРОЦЕССОРОВ

Нижегородский государственный технический университет им. Р.Е. Алексеева

Описаны базовые понятия и методы обработки цифровых сигналов, использующиеся для конструирования систем распознавания речи.

*Ключевые слова*: система распознавания речи, мел-кепстральные коэффициенты, окно Хэмминга, коэффициенты линейного предсказания, конечные точки слова, слуховой аппарат, цифровая обработка сигналов

## Введение

В современных компьютерных системах все больше внимания уделяется построению интерфейса естественного ввода-вывода информации. Одним из перспективных направлений на сегодняшний день является использование систем речевого диалога, которая предполагает автоматический синтез и распознавание речи. Подобная система может быть встроена в различные приложения, например в системы голосового контроля, голосового доступа к информационным ресурсам, обучения языку с помощью компьютера, помощи недееспособным, доступа к чему-либо через системы голосовой верификации/идентификации.

При разработке системы автоматического распознавания речи, представляющей собой наиболее сложную подсистему речевого диалога, используют различные методы обработки информации. В настоящей статье собран основной теоретический материал по данной теме и описаны подходы, которые могут использоваться для создания базовой системы распознавания речи, на основе которой могут быть построены более сложные решения.

## Основные режимы работы систем распознавания речи

Системы распознавания речи обычно имеют два режима: режим обучения и режим распознавания. Эти режимы используют общую функциональную часть (рис. 1), задача которой заключается в получении сигнала, предобработке фреймов, нахождении конечных точек слова и экстракции характеристик сигнала.

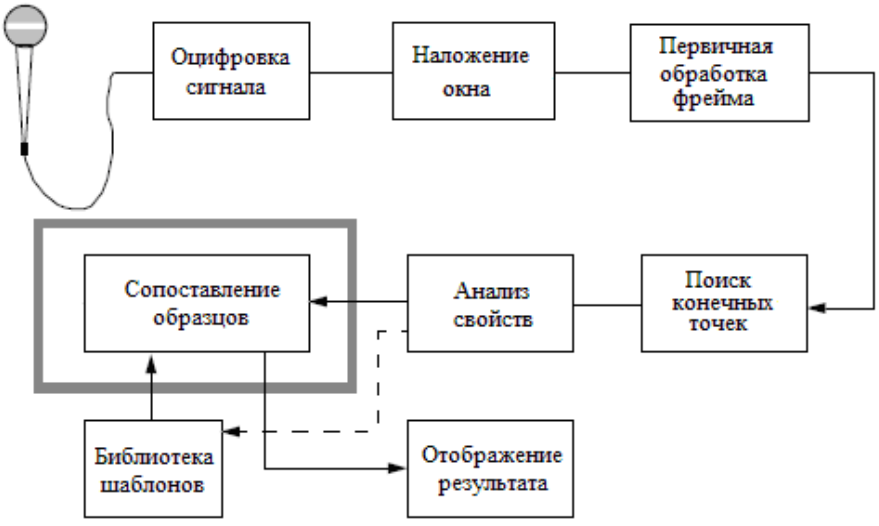



**Рис. 1. Блок-схема системы распознавания речи**

Дальнейшее поведение системы зависит от режима работы. Если система находится в режиме обучения, полученные на этапе выделения характеристик значения сохраняются в библиотеке шаблонов. При нахождении системы в состоянии распознавания, полученный набор значений сравнивается с наборами из библиотеки. Наилучший результат сравнения возвращается в качестве результата распознавания.

**Подходы к решению задачи**

Процесс решения задачи распознавания изолированных слов можно разделить на четыре основных этапа:

- ввод сигнала из внешней среды в систему;
- нахождение конечных точек слова;
- экстракция характеристик сигнала;

определение результата распознавания.

Выделение слова из непрерывного потока входящей информации является сложной задачей в силу особенностей голоса, окружающей среды и аппаратуры, с помощью которой производится запись звукового сигнала. Человек может успешно распознавать речь, громкость которой меняется в очень широких пределах. Мозг человека способен отфильтровывать тихую речь от помех окружающей среды, например музыки или шума работающих приборов. В отличие от человеческого мозга, цифровая аппаратура очень чувствительна к внешним воздействиям подобного рода. Если микрофон стоит на столе, то при повороте головы или изменении положения тела, расстояние между ртом и микрофоном будет изменяться. Это приведет к изменению уровня выходного сигнала микрофона и соотношения сигнал/шум, что, в свою очередь, ухудшит надежность распознавания речи. Изменение интенсивности речи в процессе произношения, смягчение начальных и конечных звуков слова на практике приводят к тому, что конечные точки трудноотличимы от сигнала помехи, постоянно присутствующей в сигнале.

Частота оцифровки также имеет немаловажное значение. Человеческое ухо воспринимает звук в диапазоне частот от 16 Гц до 22 кГц. В действительности, границы слышимых частот зависят от конкретного человека – его возраста, пола и здоровья. По теореме Котельникова, аналоговый сигнал может быть восстановлен однозначно и без потерь по своим дискретным отсчетам, взятым с частотой строго большей удвоенной верхней частоты. Человеческому голосу соответствует диапазон частот 300–4000 Гц, следовательно, для того чтобы восстановить голосовой сигнал без потерь, необходимо использовать частоту дискретизации, большую 8 кГц. Оптимальное значение частоты в условиях рассматриваемой задачи – 12 кГц.

Проблема изменения интенсивности произношения заключается в том, что оцифрованный сигнал имеет различную амплитуду для одного и того же слова в зависимости от изменения расстояния между диктором и микрофоном, положения источника сигнала в пространстве, а также громкости издаваемого звука. Решением в данном случае является использование автоматической регулировки усиления (АРУ). Суть данного процесса заключается в автоматическом поддержании выходного сигнала постоянным по некоторому параметру (например, амплитуде простого сигнала или мощности сложного сигнала), независимо от амплитуды (мощности) входного сигнала (рис. 2). В используемом в разработке аппаратном обеспечении USB Stick eZdsp vc5505 присутствует блок обрабоки аудиоданных на основе микросхемы AIC3204. Данный кристалл имеет встроенную функцию автоматической регулировки усиления (АРУ), что позволяет без дополнительных задержек получить оцифрованный нормированный сигнал высокого качества, который можно использовать без дополнительной обработки. Автоматическая регулировка усиления также может быть реализована программно, но скорость работы системы будет значительно ниже.

Основной единицей обработки оцифрованного сигнала является фрейм – массив отсчетов, соответствующий определенному временному промежутку. Речь является нестационарным сигналом, характеристики которого часто меняются во времени, но известно, что для

большинства фонем, характеристики сигнала остаются постоянными в течение короткого промежутка времени (~5–100 мс) и в его пределах сигнал можно считать стационарным. Цифровые сигнальные процессоры имеют программную, а некоторые реализации и аппаратную поддержку специфических для обработки сигналов математических функций, таких как преобразование Фурье. И в том, и в другом случае размер массива входных параметров для этих функций должен быть кратен $2^n$, поэтому для удобства и экономии вычислительных ресурсов размер фрейма также должен быть кратен этому значению. Оптимальным в данной реализации системы был выбран размер фрейма, равный 512 отсчетам и соответствующий ~43 мс, что удовлетворяет условию стационарности характеристик. Размер буфера для хранения оцифрованного сигнала кратен размеру фрейма и равен 12288, что соответствует 24 фреймам или ~1 с. Этого времени вполне достаточно для произношения коротких голосовых команд.

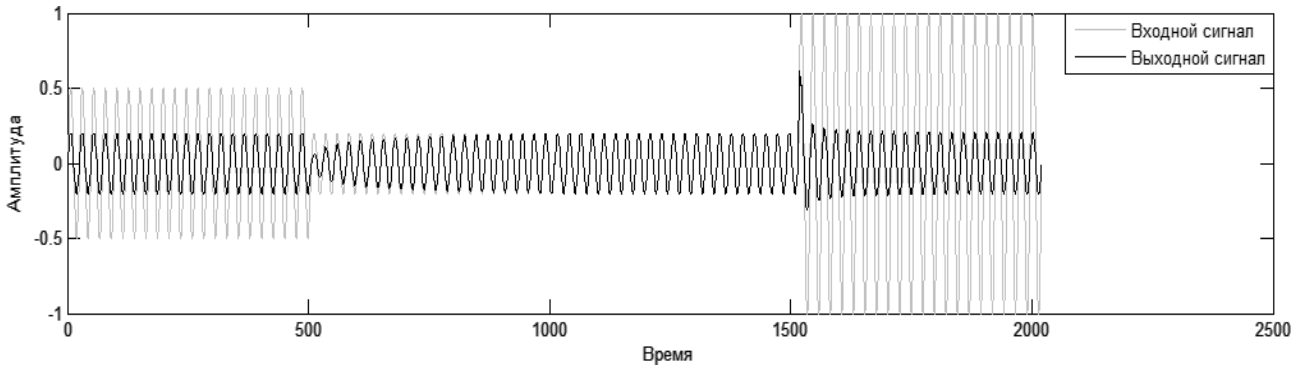



**Рис. 2. Автоматическая регулировка усиления**

## Определение конечных точек слова

Задача определения моментов начала и окончания фразы является одной из важных задач в области обработки речи. Методы обнаружения конечных точек слова используются для отделения речи от окружающего шума, а также уменьшения числа арифметических операций, поскольку обрабатываются только те сегменты, в которых имеется речевой сигнал. Проблема отделения речи от помехи очень сложна, за исключением случаев очень большого отношения сигнал/шум, т.е. в случае высококачественных записей, выполненных в студийных условиях. В этом случае энергия даже наиболее слабых звуков превышает энергию шума и, таким образом, достаточно лишь измерить энергию сигнала. Но подобные условия записи, как правило, не встречаются в реальных условиях.

Для выделения слова из непрерывного потока информации в реальном масштабе времени может использоваться простой, но в то же время достаточно эффективный метод определения конечных точек Рабинера-Самбура, основанный на подсчете энергии фрейма и частоты переходов через нуль. Данный метод требует меньшего объема вычислений из-за отсутствия дополнительного преобразования сигнала из временной области в частотную.

Под энергией фрейма в данном случае понимается нормированная сумма абсолютных значений амплитуд дискретных отсчетов сигнала (1).

$$E = \frac{1}{K}\sum_{n=1}^{N} A_n, \qquad (1)$$

где $K$ – коэффициент нормировки, $N$ – длина фрейма.

Для вычисления значения энергии могут быть использованы иные методы расчета, например нахождение евклидовой нормы. Поскольку для сигнальных процессоров серии tms320vc5505, минимальной единицей памяти является 2 байта, а также тот факт, что разрешающая способность аудио-кодека, с помощью которого производится оцифровка звукового сигнала, равна 16 битам, то в качестве структур для хранения информации предпочтение отдается массивам, содержащим двухбайтовые элементы. Нормирование конечного значения необходимо для

того, чтобы избежать перегрузки разрядной сетки. Коэффициент нормировки выбирается из следующих соображений: поскольку разрешающая способность кодека равна 16 битам, а значение амплитуды может быть как положительным, так и отрицательным, то максимально возможная по абсолютному значению величина, которую можно сохранить в двухбайтовом знаковом типе, равна $2^{16-1} = 32768$, а максимальная сумма абсолютных значений амплитуд равна 32768 * 512. Исходя из изложенного, а также того, что значение энергии хранится в двухбайтовом знаковом типе, коэффициент нормировки выбирается равным длине фрейма.

Частота переходов через нуль определяется как число раз, когда исходный сигнал меняет свой знак и его значение находится выше порога шума. Данная величина не нуждается в нормировке, поскольку максимальное значение параметра равно $N$-1.

На рис. 3 представлена временная диаграмма слова «раз» с наличием постоянной помехи. Поясним суть модифицированного метода Рабинера-Самбура для определения моментов начала и окончания слова с помощью рис. 4, *а*, *б*.

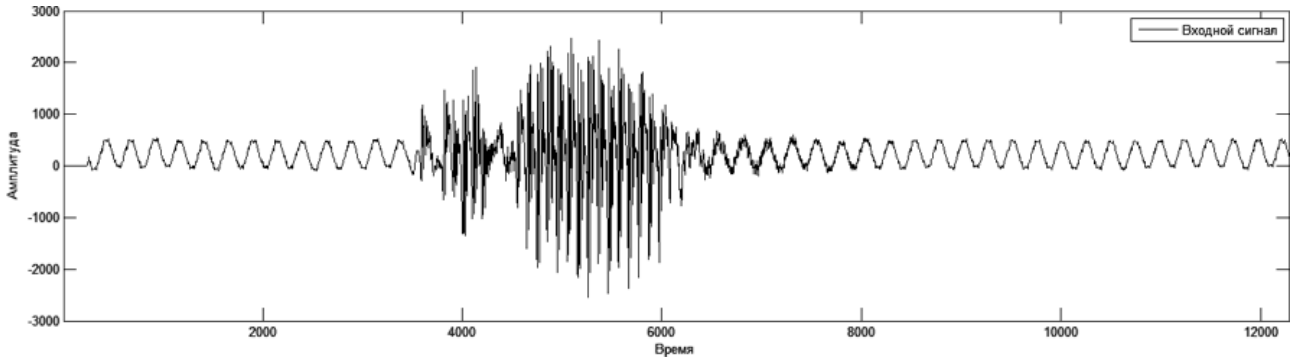
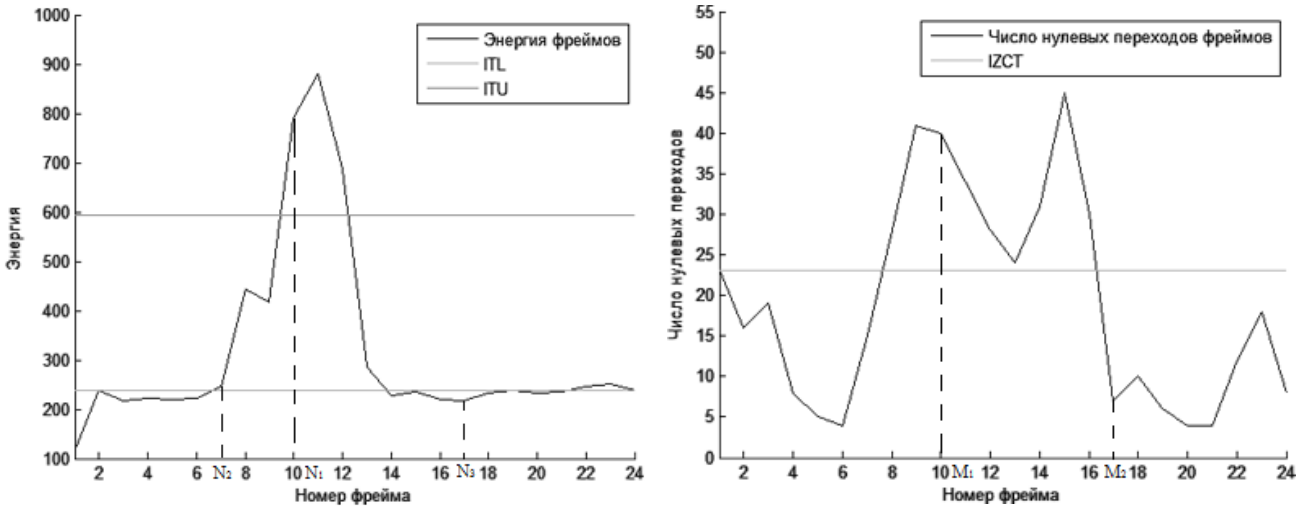



**Рис. 3. Временная диаграмма слова «раз»**




***а*** ***б***

**Рис. 4. Изменение значения:**
*а* – величины энергии фреймов слова «раз»; *б* – нулевых переходов фреймов слова «раз»

Ввод сигнала в систему начинается после нажатия пользователем кнопки записи. Предполагается, что первые два фрейма, соответствующие ~86 мс не должны содержать в себе полезной информации. По этому участку определяются статистические характеристики шума – порог для среднего числа нулевых переходов (IZCT на рис. 4, *б*) и верхний и нижний пороги энергии сигнала (на рис. 4, *а* - ITU и ITL соответственно). После этого производится поиск фрейма, в котором величина средней энергии и числа переходов через нуль превышают соответствующие пороги ITU и IZCT (1.2). Если такой фрагмент найден (значения $N_1$ на

рис. 4, *а* и $M_1$ на рис. 4, *б*) – это означает, что данный фрейм точно содержит полезный сигнал, и предполагается, что начало слова находится вне его пределов. Фреймом, содержащим начальную точку, считается тот фрагмент, для которого величина энергии впервые стала больше уровня ITL, если двигаться в обратном направлении следования блоков (соответствует значению $N_2$ – рис. 4, *а*).

Для систем реального времени, где выполнение обратных шагов не желательно, данный алгоритм может быть модифицирован для обеспечения его работы строго в прямом направлении. В данном случае сначала происходит поиск точки $N_2$, после чего ищется фрагмент, где выполняется условие (2). Если в процессе анализа сигнала встречается несколько положительных переходов через уровень ITL, предполагаемым началом слова считается фрейм, соответствующий последнему переходу.

$$\begin{cases} E_i > ITU \\ Z_i > IZCT \end{cases} \quad (2)$$

Значение величины ITU определено экспериментально и равно величине, в 2.5 раза большей уровня ITL.

После того, как начальный момент фразы определен (фрейм №7) и началась запись сигнала в память, параллельно с передачей данных происходит определение конечного момента слова. Принцип поиска схож с тем, как определяется начало слова, считается, что фрейм содержит конечную точку тогда, когда выполнено условие (3).

$$\begin{cases} E_i < ITU \\ Z_i < IZCT \end{cases}. \quad (3)$$

После нахождения фрагмента, параметры которого соответствуют данному критерию, предполагается, что дальнейший сигнал не содержит полезной информации и сравним с фоновым шумом. Для сигнала, представленного на рис. 3, конечной точкой фразы считается фрейм №17, поскольку его энергия ниже уровня ITL (точка $N_3$ на рис. 4, *а*) и число нулевых переходов впервые становится ниже уровня IZCT ($M_2$ на рис. 4, *б*).

Выполнение одного лишь условия (2) не гарантирует точного определения конечной точки. Существуют слова, содержащие периоды тишины между фонемами, например, в слове «четыре» (рис. 5) между звуками «ч» и «т» присутствует промежуток провала характеристик, который можно принять за конец слова (диапазон [9:12] на рис. 6, *а*, *б*).

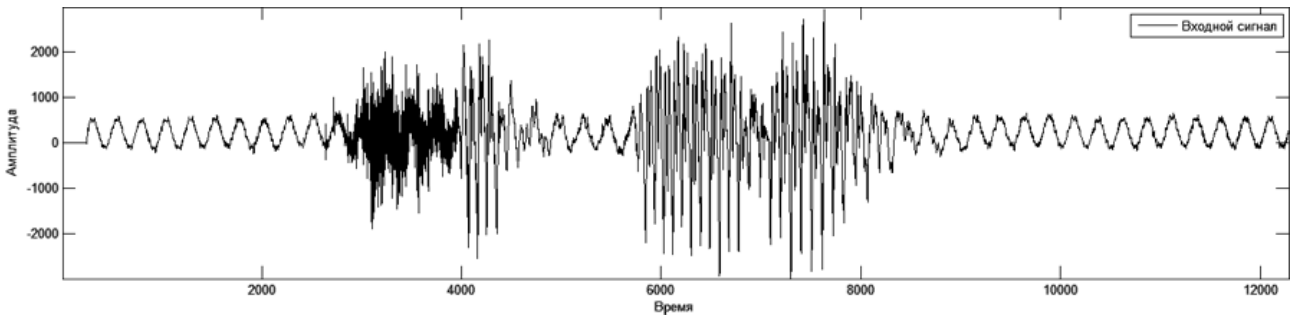
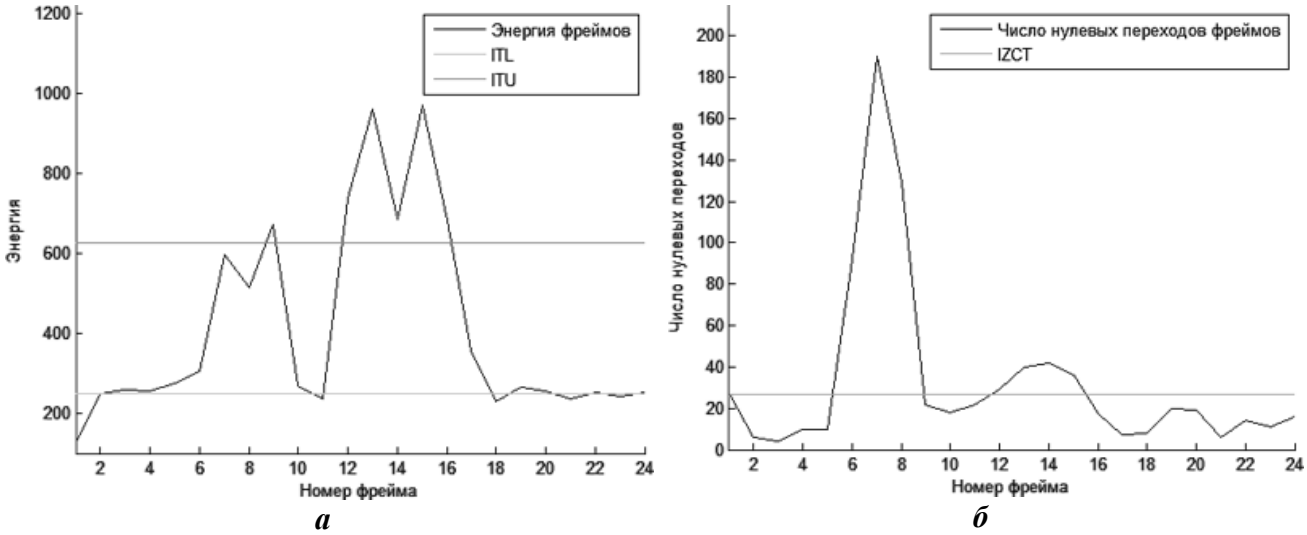



**Рис. 5. Временная диаграмма слова «четыре»**

Подобная проблема решается путем ввода значения максимальной длительности тишины – отрезка времени, в пределах которого значения параметров снова могут превысить соответствующие уровни. Если по истечении данного промежутка времени характеристики сигнала остались на уровне шума, то считается, что конец слова найден и находится в том фрагменте, в котором значение энергии впервые стало меньше уровня ITL. Обычно промежуток тишины выбирается равным ~0.1 с. В текущей реализации он равен трём фреймам, что соответствует значению времени ~130 мс.

После нахождения начального и конечного фреймов возможно проведение дополнительного анализа выделенного сигнала для определения конечных точек на уровне отсчетов.




*а* *б*

**Рис. 6. Изменение значения:**
*а* – величины энергии фреймов слова «четыре»; *б* – нулевых переходов фреймов слова «четыре»

## Выделение характеристик

Как только слово было выделено из потока входных данных, начинается следующий этап процесса распознавания – выделение характеристик. Здесь могут применяться различные методики, например методика нахождения мел-кепстральных коэффициентов или коэффициентов линейного предсказания. Главной задачей на данном этапе является выделение неких параметров сигнала, причем число этих параметров должно быть минимально, чтобы ускорить сравнение с наборами параметров из библиотеки, и в то же время данные параметры должны быть такими, чтобы по ним можно было достаточно точно определить конкретное слово.

## Мел-кепстральные коэффициенты

Эволюция сенсорных систем, которыми обладают живые существа, шла по пути: «различать, чтобы выжить». Слуховой аппарат человека как сенсорный анализатор обеспечивает различение звуков по их частотному составу. Однако реакция на звуковой стимул должна быть быстрой, а значит, обработка сигналов в ухе и нервной системе должна выполняться за небольшое время. Требования высокой частотной и временной различительной способности анализатора противоречивы, но в результатом эволюции было оптимальное сочетание этих показателей.

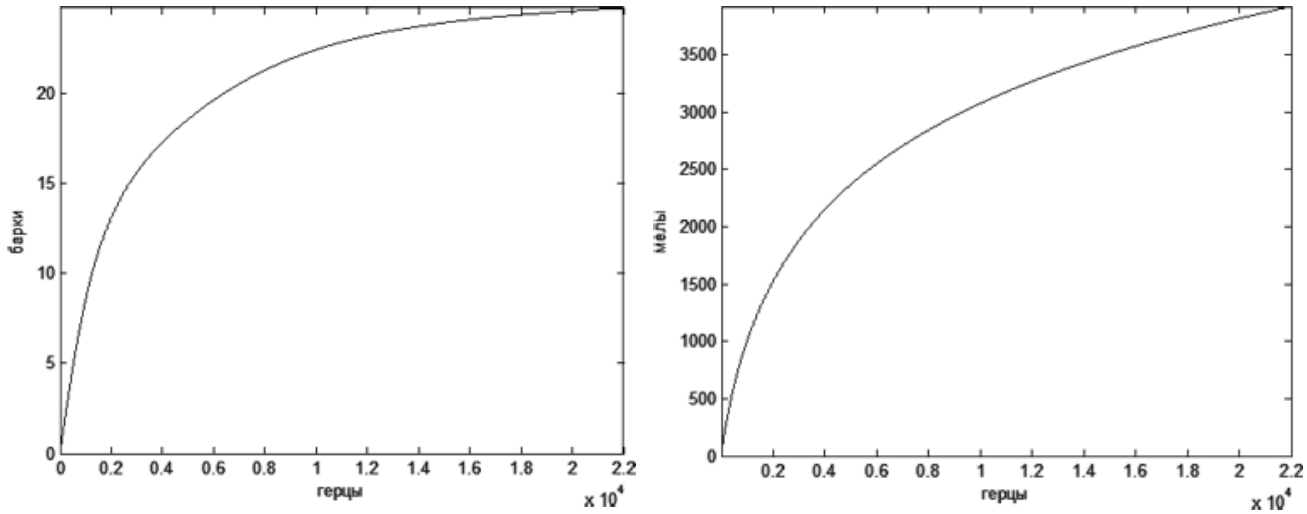
Органы слуха человека обладают свойством частотного маскирования – ситуация, когда нормально слышимый звук накрывается другим громким звуком с близкой частотой. Данная характеристика зависит от частоты сигнала и варьируется от 100 Гц для низких слышимых частот до более 4000 Гц для высоких частот.

Следовательно, область слышимых частот можно разделить на несколько критических полос (принято деление на 24 критические полосы), которые обозначают падение чувствительности уха для более высоких частот. Можно считать критические полосы еще одной характеристикой звука, подобной его частоте. Однако, в отличие от частоты, которая абсолютна и не зависит от органов слуха, критические полосы определяются в соответствии со слуховым восприятием. В итоге они образуют некоторые меры восприятия частот, для которых введены единицы измерения – барк и мел.

Шкала барков (рис. 7, *а*) связана с критическими полосами слуха, и поскольку ширина этих полос неравномерна, увеличивается с возрастанием частоты звуковых колебаний, то и сама является неравномерной. Прямая и обратная зависимости между высотой звука в барках и частотой тона в Гц определяется формулами (4) и (5) соответственно:

$$b = 13\,\mathrm{atan}(0.00076 * f) + 3.5\,\mathrm{atan}\left(\frac{f}{7500}\right)^2 \quad (4)$$

$$f = \frac{52548}{b^2 - 52.56b + 690.39} \quad (5)$$




*а* *б*

**Рис. 7. Для слышимого диапазона частот**
*а* – барк-шкала; *б* – мел-шкала

Неравномерной также является и шкала мелов (рис. 7, *б*), основанная на статистической обработке большого количества данных о субъективном восприятии высоты звуковых тонов. Результаты исследований показывают, что высота звука связана главным образом с частотой колебаний, но зависит также от уровня громкости звука и его тембра. Прямая и обратная зависимости между высотой звука в мелах и частотой тона в Гц определяется формулами (6) и (7) соответственно:

$$m = 1127.01048 \ln\left(1 + \frac{f}{700}\right), \quad (6)$$

$$f = 700\left(e^{\frac{m}{1127.01048}} - 1\right) \quad (7)$$

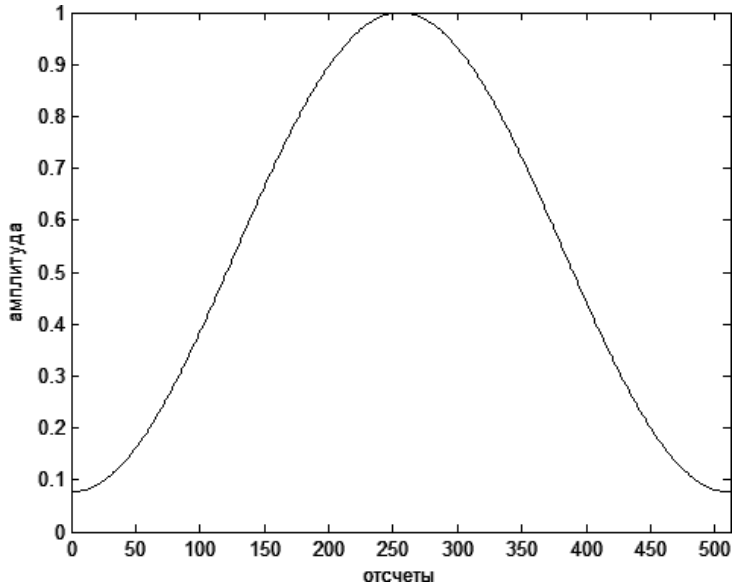



**Рис. 8. Окно Хэмминга**

Как уже было отмечено, частотный диапазон человеческого голоса весьма ограничен и располагается в интервале от 300 до 4000 Гц. Из данного факта следует, что путем моделирования полосового фильтра можно отбросить частотные составляющие, которые находятся за пределами этого диапазона и, соответственно, не несут смысловой нагрузки.

Исследуемый сигнал разделяется на фреймы на основе метода периодограмм Уэлча – вектор отсчетов сигнала делится на перекрывающиеся сегменты (как правило, используется 50%-ное перекрытие), после чего каждый фрейм умножается на весовую функцию и для него вычисляется дискретное преобразование Фурье (8). В качестве весовой функции выбрано окно Хамминга (9), представленное на рис. 8, но могут быть использованы и другие виды окон, например, окно Ханна. Применение весовой функции позволяет ослабить растекание спектра на стыках фреймов.

$$F_m = \sum_{n=1}^{N} f_n \cdot w_{n-m} \cdot e^{-j\omega n} \tag{8}$$

$$w_n = 0.53836 - 0.46164 \cos\frac{2\pi n}{N-1} \tag{9}$$

Исследования в области психофизического восприятия показали, что основная значимая информация содержится в действительном частотном спектре, поэтому после выполнения преобразования Фурье для дальнейшего анализа выделяется действительный спектр сигнала, а информация о фазе может быть опущена. При помощи смоделированного полосового фильтра отбрасывается информация о частотных составляющих, не находящихся в диапазоне [300, 4000] Гц. Затем на полученный фрейм накладываются взвешенные треугольной функцией перекрывающиеся окна, у которых значения центральных частот изменяются нелинейно в соответствии с мел-шкалой (рис. 9). Величины центральных частот вычисляются следующим образом:

- задается количество мел-фильтров и границы диапазона частот, для которого будет производится фильтрование;
- в соответствии с формулой (6) выполняется преобразование границ диапазона из Гц в мел и вычисляется шаг изменения частоты n по формуле (10);
- для каждого значения из диапазона [$m_{max}$, $m_{min}$] с шагом n выполняется обратное преобразование из мел в Гц (1.7).

$$n = \frac{m_{\max} - m_{\min}}{k} \tag{10}$$

где $m_{max}$, $m_{min}$ – границы диапазона в мелах; $k$ – количество мел-фильтров.

В пределах полученных окон вычисляются средние значения действительного спектра, в результате чего получается сглаженный сильно кореллированный мел-спектр с различной детализацией диапазонов частот психофизической модели звукового восприятия. На рис. 10, *а, б* изображен спектр случайного сигнала и его сглаженное мел-представление.

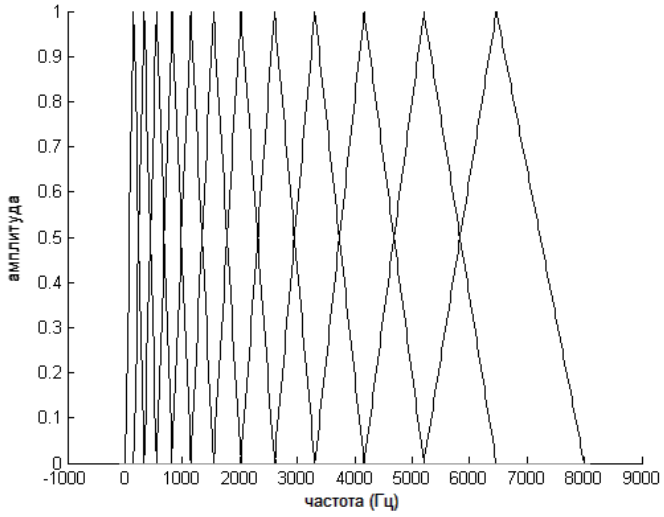
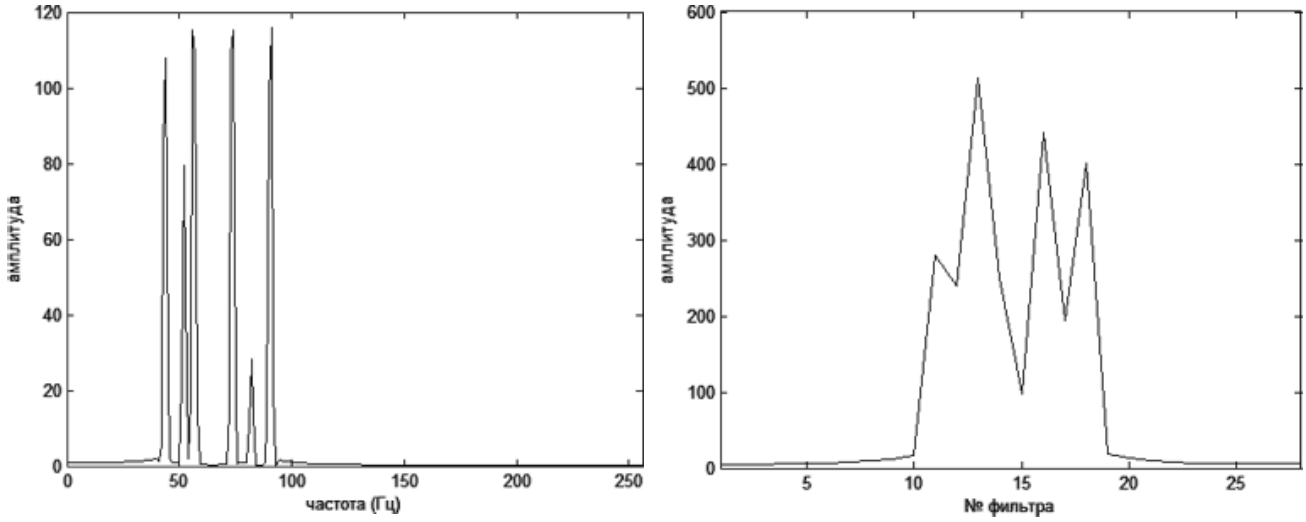



**Рис. 9. Применение банка мел-фильтров**




*а* *б*

**Рис. 10. Спектр сигнала:**
*а* – частотный; б – мел-спектр сигнала для банка из 28 фильтров

Для уменьшения количества выходных параметров и декорреляции компонентов выполняется последний шаг вычисления мел-кепстральных коэффициентов. Для этого в литературе предлагается использовать метод главных компонент, который иногда называют преобразованием Карунена-Лоэва или методом Хотеллинга. В области обработки речевых сигналов для уменьшения вычислений можно использовать дискретно-косинусное преобразование (11), которое дает схожие результаты.

$$X_k = \alpha_k \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], \quad k = \overline{0, N-1}, \tag{11}$$

где $$\alpha_k = \begin{cases} \sqrt{\frac{1}{N}}, \text{при } k = 0 \\ \sqrt{\frac{2}{N}}, \text{при } k \neq 0 \end{cases}$$

Полученный набор значений называется мел-кепстральными коэффициентами. Обычно сохраняется только первые несколько элементов (от 8 до 16), по которым в дальнейшем производится идентификация сигнала.

### Коэффициенты линейного предсказания

Кодирование с линейным предсказанием – это техника кодирования речевой информации путем моделирования голосового тракта. В соответствии с моделью, человек может производить два типа звуков: вокализованные и невокализованные. Если голосовые связки вибрируют при прохождении через них воздуха из легких, то производимые при этом звуки называются вокализованными. Звуки, полученные при участии только языка, зубов и губ, называются невокализованными.

Для обоих типов звуков голосовой тракт может быть представлен как последовательность цилиндров различных радиусов с различным значением энергии на границах между цилиндрами. Математически данную модель можно представить в виде линейного фильтра, возбуждаемого основной частотой для представления вокализованных звуков, либо белым шумом для невокализованных.

Задача анализа на основе линейного предсказания – получить параметры, необходимые для воссоздания исходного звука: тип звука, значение основной частоты, коэффициенты фильтра. Цель кодирования с линейным предсказанием заключается в моделировании сигнала как линейной комбинации предыдущих отсчетов (12). Данный метод является достаточно эффективным, поскольку речь – это высоко коррелированный в течение короткого промежутка времени сигнал, а значит предсказание может быть выполнено с минимальной ошибкой. В задаче распознавания речи данный метод применяется для моделирования спектра сигнала как авторегрессионного процесса.

$$s(n) = -\sum_{i=1}^{N_{lp}} a_{N_{lp}}(i) * s(n-i) + e_n, \tag{12}$$

где $N_{lp}$ – порядок предсказания (количество коэффициентов модели); $a_{N_{lp}}$ – коэффициенты линейного предсказания; $e_n$ – функция ошибки модели (различие между предсказанным и реальным значениями).

Учитывая то, что квадратичная ошибка должна быть минимальна, коэффициенты линейного предсказания определяются из следующей системы нормальных уравнений, представленной в матричном виде:

$$R_{N_{lp}} a_{N_{lp}} = -r_{N_{lp}} \tag{13}$$

где

$$R_{N_{lp}} = \begin{bmatrix} r_0 & r_1 & r_2 & \cdots & r_{N_{lp}-1} \\ r_1 & r_0 & r_1 & \cdots & r_{N_{lp}-2} \\ r_2 & r_1 & r_0 & \cdots & r_{N_{lp}-3} \\ \cdots & \cdots & \cdots & & \cdots \\ r_{N_{lp}-1} & r_{N_{lp}-2} & \cdots & & 0 \end{bmatrix}$$

$$a_{N_{lp}} = \begin{bmatrix} a_1(1) & \cdots & a_{N_{lp}}(N_{lp}) \end{bmatrix}^T$$

$$r_{N_{lp}} = [r_1 \quad \cdots \quad r_{N_{lp}}]^T$$

где $r_k$ – $k$-й коэффициент автокорреляции речевого сигнала, взвешенного оконной функцией $w$ (обычно используют окно Хамминга) (14).

$$r_k = \sum_{n=k}^{N-1} w(n) * s(n) * w(n-k) * s(n-k). \tag{14}$$

Коэффициенты линейного предсказания вычисляются следующим образом:

$$a_{N_{lp}} = -R_{N_{lp}}^{-1} r_{N_{lp}}. \tag{15}$$

Матрица автокорреляции $R_{N_{lp}}$ имеет структуру Теплица, для решения которой существует эффективный алгоритм Левинсона-Дарбина, суть которого можно кратко представить в псевдокоде:

$$e_0 = r_0$$

$$for\ 1 \le m \le N_{lp}$$

{

$$a_m(0) = 1$$

$$a_m(m) = k_m = \frac{-r_m - \sum_{i=1}^{m-1} a_{m-1}(i) * r_{m-1}}{e_{m-1}}$$

$$a_m(j) = a_{m-1}(j) + k_m * a_{m-1}(m-j), \quad j = \overline{1, m-1}$$

$$e_m = e_{m-1} * (1 - k_m^2)$$

}

Для уменьшения количества сохраняемых параметров к полученным на предыдущем шаге коэффициентам применяется преобразование (16) для расчета кепстральных коэффициентов.

$$c_i = \left[DCT\left(ln(a_i{}^2)\right)\right]^2 \quad (16)$$

где DCT – дискретно-косинусное преобразование (11).

**Динамическое выравнивание времени**

Заключительным этапом распознавания является сопоставление входного образа с набором эталонных образов из библиотеки. Результатом распознавания является индекс библиотечного шаблона, который имеет наибольшее сходство с исходным блоком. Но различные реализации речевых образов, относящихся к одному и тому же классу, могут значительно отличаться друг от друга по длительности: это связано с нестабильностью темпа речи диктора, вызванной влиянием интонации, акцента и т.п. Для корректного сопоставления речевых образов необходимо производить их выравнивание по длине. Выравнивание путём линейного сжатия или растяжения одной реализации слова до величины другой не решает данную задачу, поскольку речевой сигнал протекает во времени неравномерно. Это свойство речи выражается в неравномерном изменении длительности звуков слова при изменении длительности слова в целом, поэтому сопоставление целесообразно выполнять с помощью нелинейной временной нормализации.

Для нелинейного выравнивания сопоставляемых образов используется алгоритм, основанный на определении наилучшего соответствия входных и эталонных речевых образов, известный как метод динамического выравнивания времени (dynamic time warping). Суть алгоритма заключается в следующем. Обозначим расстояние между *i*-м элементом массива параметров входного образа и *j*-м элементом массива параметров эталона как $D_{ij}$. Для нахождения элементов входного вектора, наилучшим образом соответствующих элементам эталона, определяется матрица C размера *M*N* по следующим формулам:

$$\begin{aligned} &\mathrm{C}_{1,1} = D_{1,1} \\ &C_{i,1} = D_{\mathrm{i},1} + \mathrm{C}_{\mathrm{i}-1,1}, \qquad i = \overline{2..M} \\ &C_{1,j} = D_{1,\mathrm{j}} + C_{1,j-1}, \qquad j = \overline{2..N} \\ &C_{i,j} = D_{i,j} + \min\left[C_{i-1,j},\ C_{i-1,j-1},\ C_{i,j-1}\right], \qquad i = \overline{2..M}, j = \overline{2..N} \end{aligned} \quad (17)$$

где *M* – количество элементов входного образа; *N* – количество элементов эталона.

Расстояние $D_{ij}$ может вычисляться различными способами, например, как евклидово расстояние (18), манхеттенское расстояние (19) или расстояние Итакуры-Саито (20). Последнее используется в случае, если вектор характеристик содержит коэффициенты линейного предсказания.

$$D_{ij} = \sqrt{x_i^2 + x_j^2} \quad (18)$$

$$D_{ij} = \left|x_i - x_j\right| \quad . \quad (19)$$

$$D_{ij} = \frac{x_j}{x_i} - \ln\left(\frac{x_j}{x_i}\right) - 1 \quad (20)$$

На рис. 11 ломаной линией соединены элементы матрицы C, соответствующие наиболее схожим элементам входного слова и эталона. Вертикальный отрезок представляет случай, когда несколько элементов эталона соответствует одному элементу входного вектора. Горизонтальному отрезку соответствует случай, когда несколько элементов массива входных параметров соответствуют одному элементу эталона. Элемент $\mathrm{C}_{\mathrm{M,N}}$ содержит суммарную оценку схожести двух векторов характеристик. После сравнения параметров входного слова со всеми шаблонами из библиотеки среди полученных суммарных оценок выбирается минимальная, а индекс соответствующего ей шаблона выводится в качестве результата распознавания.

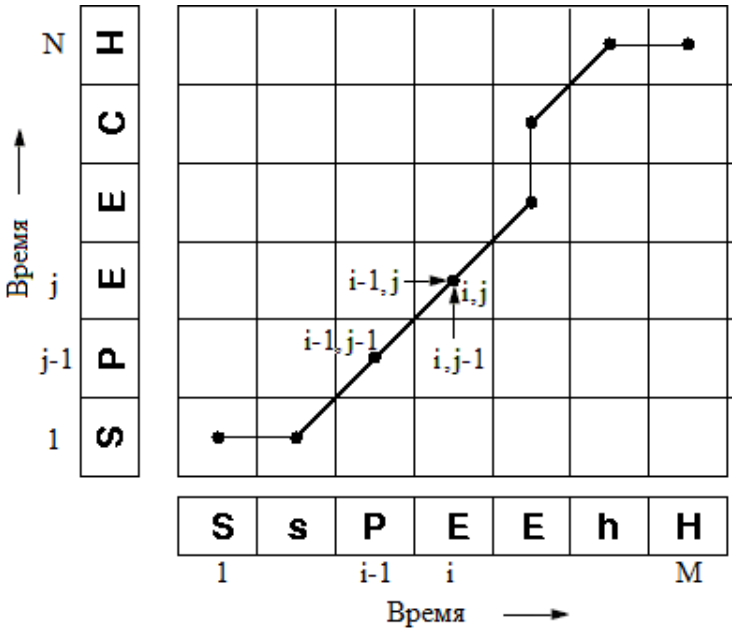



**Рис. 11. Динамическое выравнивание времени**

Рассмотренные в статье методы обработки сигналов были реализованы в виде библиотеки функций C++ для использования на цифровых сигнальных процессорах tms320vc5505 марки Texas Instruments. На основе данной библиотеки была создана тестовая дикторозависимая система распознавания изолированных слов с маленьким размером словаря.

**Библиографический список**

1. **Бондаренко, И.Ю.** Анализ эффективности метода нечёткого сопоставления образов для распознавания изолированных слов / И.Ю. Бондаренко, О.И. Федяев // Интеллектуальный анализ информации ИАИ-2006: сб. трудов VI междунар. науч. конференции; под ред. Т.А. Таран. – К.: Просвіта, 2006. С. 20–27.
2. Рабинер Л.Р., Цифровая обработка речевых сигналов: [пер. с англ.] / Л.Р. Рабинер, Р.В. Шафер. – М.: Радио и связь, 1981. – 251 с.
3. **Сэломон, Д.** Сжатие данных, изображений и звука / Д. Сэломон. – М.: Техносфера, 2004. – 368 с.
4. Beth Logan, Mel frequency cepstral coefficients for music modeling [Электронный ресурс]. // Cambridge research laboratory, Compaq computer corporation. - Режим доступа: http://apotheca.hpl.hp.com/ftp/pub/compaq/CRL/publications/logan/musicir_paper.pdf
5. Joseph Picone, Fundamentals of speech recognition [Электронный ресурс]. // Mississippi state university, Department of electrical and computer engineering. - Режим доступа: http://www.isip.piconepress.com/publications/courses/msstate/ece_8463/lectures/current/
6. Cedrick Collomb, Linear Prediction and Levinson-Durbin Algorithm / Cedrick Collomb. [Электронный ресурс]. // - Режим доступа: http://www.emptyloop.com/technotes/A%20tutorial%20on%20linear%20prediction%20and%20Levinson-Durbin.pdf
7. Mel scale - Wikipedia, the free encyclopedia [Электронный ресурс]. // - Режим доступа: http://en.wikipedia.org/wiki/Mel_scale, свободный. – Загл. с экрана.
8. Bark scale - Wikipedia, the free encyclopedia [Электронный ресурс]. // - Режим доступа: http://en.wikipedia.org/wiki/Bark_scale, свободный. – Загл. с экрана.

*Дата поступления*
*в редакцию 10.10.2012*

**O. Agashin, O. Korelin**

# GENERAL METHODS OF DIGITAL PROCESSING OF THE SPEECH SIGNAL IN CASE OF ISOLATED WORDS RECOGNITION ROBLEM WITH DSP APPLICATION

Nizhny Novgorod state technical university n.a. R.E. Alexeev

**Purpose** Development and deployment the real-time speech recognition systems for controlling industrial or home electronics. Providing fast and secure solution based on energy effective and independent of any operating system hardware with small form factor.

**Approach** A theoretical framework is proposed combine distinct techniques and methods preferred for speech recognition. Optimization of finding end points and decision making algorithms provides the most efficient solution for interactive control of small systems.

**Research limitations/implications** The present study provides a starting-point for further research on the speech recognition problem. For example, Hidden Markov model, wavelets and neuron network could be involved for making more precise prediction.

**Value** Besides, new software library providing API for processing of digital signals developed during the research. It supplies the base functionality necessary for building a simple real-time recognition system coupled with DSP-based hardware.

*Key words*: isolated words, speech recognition, digital signal processing, search endpoints, hardware based solution.