МЕТОДИКА РАСПОЗНАВАНИЯ РЕЧИ

Федоров Е.Е.
Донецкий национальный университет, г.Донецк
кафедра прикладной математики и теории систем управления
E-mail: fee@iai.donetsk.ua

Abstract
Fedorov U. Technique of speech recognition. *For development of the natural - language interface of the automated system control (ASC) in clause the technique of recognition of speech used for recognition of the commands, sent by the operator is offered. For a choice of an effective system of features the numerical research was carried out.*

Постановка проблемы. В настоящее время актуальной является разработка систем, предназначенных для распознавания речи. Эти системы имеют широкую область применения – ЕЯ-интерфейсы информационных, поисковых, экспертных систем и др. Применительно к ЕЯ-интерфейсу АСУ они могут использоваться для распознавания поданных оператором команд. При разработке таких систем важную роль играет выбор системы признаков.

Анализ исследований. В работах [1-4] приведены системы распознавания речи, дающие в большинстве случаев вероятность распознавания ниже 90%.

Постановка задачи. *Целью настоящей работы* является создание методики распознавания речи, базирующейся на эффективной системе признаков и использующим ее методе распознавания.

Решение задачи. В статье для методики распознавания речи рассматриваются и численно исследуются системы признаков, основанные на:

- дискретном преобразовании Фурье;
- непрерывном и дискретном вейвлет-преобразовании;
- кепстральном анализе;
- линейном предсказании;
- нормированном количестве импульсов равной длины,

и связанный с этими признаками метод распознавания, основанный на алгоритме динамического искажения времени DTW.

1 Создание систем признаков для распознавания речи

В настоящее время существуют несколько подходов для конструирования систем признаков, используемых при распознавании речи. В статье рассматриваются основные из них.

Банк фильтров можно рассматривать как упрощенную модель человеческой слуховой системы. При этом применяемые фильтры должны строго разделить частотную область сигнала на непересекающиеся участки, соответствующие основным спектральным полосам разных классов звуков речи. После применения банка полосовых фильтров к речевому сигналу получится набор спектральных составляющих исходного сигнала, на основе которых проводится анализ речи.

Вследствие изменения свойств речевого сигнала во времени его анализ проводится на фреймах длины N:

$$\hat{s}(m) = s(m)w(m),\ w(m) = 0.54 + 0.46\cos\frac{2\pi m}{N} \qquad (1)$$

где $w(m)$ - оконная функция Хэмминга

В настоящее время вместо банка цифровых фильтров используют дискретное преобразование Фурье (ДПФ), позволяющее ускорить процесс преобразования сигнала с целью формирования эталона. Спектр дискретного речевого сигнала $s(n)$ длиной ΔN представлено в виде:

$$S(k)=\sum_{n=0}^{N-1}\hat{s}(n)\,e^{-j\frac{2\pi nk}{N}},\ 0\le k\le N/2-1 \qquad (2)$$

Однако, преобразование Фурье, традиционно применяемое для обработки речевых сигналов, имеет следующие недостатки:

1) Сложность анализа нестационарных сигналов.

2) Невозможность точного восстановления сигнала из-за эффекта Гиббса. Использование оконных функций, которые борются с этим эффектом, ухудшает восстановление сигнала на участках его быстрых изменений.

3) Отсутствие хорошей частотно-временной локализации.

Ввиду этих недостатков, в последнее время вместо преобразования Фурье получили распространение методы обработки сигнала, базирующиеся на вейвлет-преобразовании.

Дискретное вейвлет-разложение сигнала $s(n)$ на P уровней представляет собой свертку на текущем i-том уровне ($i\in\overline{1,P}$) сигнала с полосовыми фильтрами с коэффициентами g_n, h_n для получения высоко- (d_{im}) и низкочастотных (c_{im}) составляющих [5-6]:

$$d_{im}=2^{1/2}\sum_{n=0}^{N/2^{i-1}-1}c_{i-1,n}g_{n+2m},\ c_{im}=2^{1/2}\sum_{n=0}^{N/2^{i-1}-1}c_{i-1,n}h_{n+2m} \qquad (3)$$

где $c_{0n}=s(n)$, $m\in\overline{0,N/2^{i-1}-1}$

Дискретное вейвлет-преобразование формирует компактный результирующий набор коэффициентов, но при этом накладывает серьезные ограничения на выбор масштабов преобразования. Масштаб, на котором проводится анализ сигнала, может быть выбран только из фиксированного ряда значений. Непрерывное вейвлет-преобразование (4) сигнала $s(t)$ является наиболее информативным представлением частотно-временных и масштабно-временных свойств сигнала:

$$CWT_s(a,b)=|a|^{-1/2}\int_{-\infty}^{\infty}s(t)\psi\left(\frac{t-b}{a}\right)dt, \qquad (4)$$

где $\psi(t)$ – вейвлет, a - масштабный коэффициент, b – сдвиг.

Чтобы получить вейвлет-коэффициенты дискретного сигнала $s(n)$ длиной N, необходимо применить численное интегрирование и заменить интегралы в (4) суммами. В результате чего получим формулу, представляющую собой аппроксимацию непрерывного вейвлет-преобразования:

$$d_{ml}=\sum_{n=0}^{N-1}s(n)\psi_{ml}(n)\Delta t,\ j_{\min}\le m\le j_{\max},\ 0\le l\le N-1 \qquad (5)$$

где Δt – величина, обратная частоте дискретизации; $\psi_{ml}(t)=a_0^{-m/2}\psi\left(a_0^{-m}t-b_0l\right)$, $a_0>1, b\ne 0$, $j_{\max}, j_{\min}$ - максимальный и минимальный уровни разложения.

Другим подходом выделения акустических параметров, основанным на теории образования речи, является метод FB-20, основанный на кепстральном анализе и осуществляющий вычисление мел-частотных кепстральных коэффициентов (MFCC).

$$\hat{s}(m) = s(m)w(m),\ w(m) = 0.54 + 0.46\cos\frac{2\pi m}{N} \tag{6}$$

$$\hat{S}(k) = \sum_{m=0}^{N-1}\hat{s}(m)e^{-j(2\pi/N)km},\quad k \in \overline{0,N-1} \tag{7}$$

$$E_i = \lg\left(\sum_{k=k1_i}^{k2_i}\left(\hat{S}(k)\right)^2 w(k)\right),\ i \in \overline{1,P}, \tag{8}$$

$$w(k) = \begin{cases} 0, & k < k1_i \\ \dfrac{k-k1_i}{\Delta K_i/2}, & k1_i \le k \le k1_i + \Delta K_i/2 \\ \dfrac{k2_i-k}{\Delta K_i/2}, & k1_i + \Delta K_i/2 \le k \le k2_i \\ 0, & k > k2_i \end{cases}$$ - треугольное окно, $\Delta K_i = k2_i - k1_i$, (9)

$$MFCC_j = \sum_{i=1}^{P} E_i \cos\left(j(i-0.5)\pi/P\right),\ j \in \overline{1,P} \tag{10}$$

где P – количество мел-частотных полос.

К подходам выделения акустических параметров, основанных на теории образования речи, относится также метод кодирования с линейным предсказанием (КЛП).

Линейный предсказатель порядка p с коэффициентами a_k определяется как система:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k), \tag{11}$$

имеющая передаточную функцию:

$$K(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{12}$$

где G - коэффициент усиления модели.

Пусть $R(i)$ – автокорреляционная функция:

$$\hat{s}(m) = s(m)w(m),\ w(m) = 0.54 + 0.46\cos\frac{2\pi m}{\Delta N} \tag{13}$$

$$R(i) = \sum_{m=0}^{N-1-i}\hat{s}(m)\hat{s}(m+i),\ 1 \le i \le p \tag{14}$$

где i – порядок линейного предсказателя.

Коэффициенты линейного предсказания a_j, согласно алгоритму Дарбина, вычисляются следующим образом:

$$E^{(0)} := R(0) \tag{15}$$

$$k_i := \left[R(i) - \sum_{j=1}^{i-1}\alpha_j^{(i-1)}R(i-j)\right] \Big/ E^{(i-1)},\ 1 \le i \le p \tag{16}$$

$$\alpha_i^{(i)} := k_i \tag{17}$$

$$\alpha_j^{(i)} := \alpha_j^{(i-1)} - k_i\alpha_{i-j}^{(i-1)},\ 1 \le j \le i-1 \tag{18}$$

$$E^{(i)} := (1-k_i^2)E^{(i-1)} \tag{19}$$

$$a_j := \alpha_j^{(p)}, 1 \le j \le p \tag{20}$$

где $\alpha_j^{(i)}$ - j-й коэффициент линейного предсказателя порядка i; k_i - i-й коэффициент отражения; $E^{(i)}$- среднеквадратичная погрешность предсказания для линейного предсказателя порядка i.

Автокорреляционная функция $r(n)$ коэффициентов КЛП вычисляется согласно (21), при этом a_0=1.

$$r(0) = \sum_{j=0}^{p} a_j^2 \ , \ r(n) = 2\sum_{j=0}^{p-n} a_j a_{j+n} \ , \tag{21}$$

Энергетический спектр определяется как:

$$W(k) = \frac{G^2}{\left(1 - \sum_{m=1}^{p} a_m \cos\left(\frac{2\pi}{N} km\right)\right)^2 + \left(\sum_{m=1}^{p} a_m \sin\left(\frac{2\pi}{N} km\right)\right)^2}, \ k \in \overline{0, N/2-1}, \tag{22}$$

где $G^2 = R(0) - \sum_{k=1}^{p} a_k R(k)$

Для сглаживания спектра обычно берется его логарифм $10\lg W(k)$.

Другим представлением сигнала является кепстр импульсной характеристики системы линейного предсказания (12). Комплексный кепстр $\hat{h}(m)$ импульсной характеристики получается с помощью рекурсивных соотношений (23), при этом a_0=1.

$$\tilde{h}(m) = a_m - \sum_{k=1}^{m-1} (k/m)\tilde{h}(k)a_{m-k} \ , \ m \in \overline{1, p} \tag{23}$$

Кроме рассмотренных выше подходов к выбору акустических параметров для представления речевого сигнала, базирующихся на физиологических и психофизических особенностях слушателя и на акустической теории образования речи, в работе также предлагается в качестве признаков использовать нормированное количество импульсов равной длины (НКИРД).

Дискретный речевой сигнал $s(n)$ подвергается многократному сглаживанию фильтром:

$$v(m) = \frac{s(m-1) + s(m) + s(m+1)}{3}, \tag{24}$$

после чего вычисляется разность исходного и сглаженного сигналов:

$$\tilde{v}(m) = s(m) - v(m) \tag{25}$$

Для сигнала $\tilde{v}(m)$ вычисляется d_z - количество импульсов длины z.

Затем определяется НКИРД согласно (26)

$$\| d_z \| = d_z \Big/ \sum_{s=1}^{len} d_s \tag{26}$$

Исходя из вышесказанного, в работе исследуются следующие системы признаков, полученные для сигнала длиной N на основе Фурье-, вейвлет-преобразований, КЛП и НКИРД:

1. Нормированный энергетический спектр, вычисленный на основе энергетического спектра ДПФ (2):

$$X_k = \frac{S^2(k)}{\sum_{i=0}^{N/2-1} S^2(i)},\ 0 \le k \le N/2-1, \tag{27}$$

2. Мера контрастности, построенная на основе ДПФ (2) и характеризующая изменение энергии в зависимости от полосы частот:

$$X_k = \lg\left(\frac{E_{FFT}(k)}{\sum_{i=0}^{k} E_{FFT}(i)}\right),\ 1 \le k \le L, \tag{28}$$

где L – количество спектральных полос,

$E_{\mathrm{FFT}}(k) = \sum_{n=N1_k}^{N2_k} S^2(n)$ - энергия спектра на k-ой полосе, имеющей границы $N1_k$ и $N2_k$.

3. Мера контрастности, построенная на основе быстрого вейвлет-преобразования (3):

$$X_k = \lg\left(\frac{E_{DWT}(k+1)}{\sum_{j=1}^{k+1} E_{DWT}(j)}\right),\ 1 \le k \le P \tag{29}$$

где $E_{\mathrm{DWT}}(k) = \sum_{n=0}^{N-1} d_{kn}^2$ - энергия вейвлет-спектра на i-том уровне разложения.

4. Мера контрастности, построенная на основе аппроксимированного непрерывного преобразования (5):

$$X_k = \lg\left(\frac{E_{CWT}(k + j_{\min})}{\sum_{j=j_{\min}}^{k+j_{\min}} E_{CWT}(j)}\right),\ 1 \le k \le j_{\max} - j_{\min} \tag{30}$$

где $E_{\mathrm{CWT}}(i) = \sum_{n=0}^{N-1} d_{in}^2$ - энергия вейвлет-спектра на i-том уровне разложения.

5. Коэффициенты MFCC, вычисляемые с помощью алгоритма FB-20:

$$X_k = MFCC_k,\ k \in \overline{1,P} \tag{31}$$

6. Коэффициенты КЛП, вычисляемые с помощью алгоритма Дарбина:

$$X_k = a_k, k \in \overline{0, p-1} \tag{32}$$

7. Коэффициенты отражения КЛП, определяемые по алгоритму Дарбина:

$$X_i = k_i, i \in \overline{1, p} \tag{33}$$

8. Кепстр импульсной характеристики системы линейного предсказания, вычисляемый по (17):

$$X_k = \hat{h}(k), k \in \overline{1, p} \tag{34}$$

9. Площади поперечных сечений кусочно-постоянной акустической трубы, содержащей (p+1) цилиндрическую секцию фиксированной длины, вычисляемые с помощью коэффициентов отражения k_i:

$$X_{i-1} = A_i,\ i \in \overline{2, p+1}, \tag{35}$$

где $A_1 = 1,\ A_{i+1} = \dfrac{1-k_i}{1+k_i} A_i,\ i \in \overline{2, p+1}$

10. Автокорреляция КЛП:

$$X_k = r(k),\ k \in \overline{1, p} \tag{36}$$

где $r(k)$ - автокорреляция КЛП, получаемая из (21)

11. Нормированный энергетический спектр КЛП:

$$X_k = \frac{W(k)}{\sum_{i=0}^{N/2-1} W(i)},\ 0 \le k \le N/2-1, \tag{37}$$

где $W(k)$ – энергетический спектр КЛП, определяемый из (22).

12. Меры контрастности энергетического спектра КЛП:

$$X_k = \lg\left(\frac{E_{LPC}(k)}{\sum_{i=0}^{k} E_{LPC}(i)}\right),\ 1 \le k \le L, \tag{38}$$

где L – количество спектральных полос,

$E_{LPC}(k) = \sum_{n=N1_k}^{N2_k} W(n)$ - энергия спектра на k-ой полосе, имеющей границы $N1_k$ и $N2_k$.

13. Нормированная автокорреляция:

$$X_k = \frac{R(k)}{R(0)},\ k \in \overline{1, p}, \tag{39}$$

где $R(k)$– автокорреляция, вычисляемая по (14).

14. НКИРД:

$$X_k = \| d_k \|, \tag{40}$$

где $\| d_k \|$ вычисляются согласно (26).

2 Метод распознавания речи

Алгоритм DTW [7,8] обеспечивает сопоставление одинаковых слов с разным темпом произнесения (и соответственно с разной длиной). Сопоставление распознаваемого сигнала и эталона отражено на рис.1. Эталон показан вертикально, а входной сигнал горизонтально. Входной сигнал “SsPEEhH” сравнивается со всеми эталонами, хранящимися в словаре. Наиболее вероятный эталон – это такой, для которого найдено минимальное расхождение между входным сигналом и эталоном.

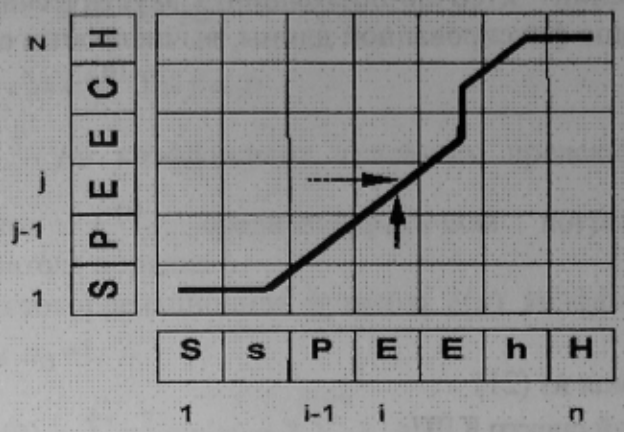

Рисунок 1 - Сопоставление эталона “SPEECH” и распознаваемого сигнала “SsPEEhH”

Алгоритм динамического искажения времени DTW используется для эффективного поиска минимального расхождения между входным сигналом и эталоном. Его ключевая идея заключается в том, что в точке (*i*,*j*) просто продолжаем самый близкий маршрут сравнения из (*i*-1,*j*-1), (*i*-1,*j*) или (*i*,*j*-1).

Пусть C_{ij} – расстояние между левыми частями распознаваемого слова (фреймы от 1 до *i*) и эталона (фреймы от 1 до *j*). D_{ij} – расстояние между *i*-м фреймом распознаваемого слова и *j*-м фреймом эталона. В качестве D_{ij} чаще всего выбирают евклидову метрику

$$D_{ij} = \sqrt{\sum_s (\tilde{X}_{is} - X_{js})^2},$$

где $\tilde{X}_{is}$ – *s*-й признак *i*-го фрейма распознаваемого слова,

X_{js} – *s*-й признак *j*-го фрейма эталона слова.

Работа алгоритма DTW состоит из следующих шагов:

1) $C_{11} = D_{11}$

2) $C_{ij} = D_{ij} + \min(C_{i-1,j}, C_{i,j-1}, C_{i-1,j-1}),\ i \in \overline{1,L}, j \in \overline{1,L}$

3) $result = C_{LL}$

В качестве эталона выбирается набор векторов признаков фреймов.

Преимуществом этого подхода является высокая точность распознавания (с ней сравнима только точность, достигаемая посредством непрерывных СММ и нейросетей). Недостатком – эталоны занимают большой объем памяти, при большом количестве эталонов процедура распознавания может быть дольше, чем у непрерывных СММ или нейросетей.

3 Численное исследование по распознаванию речи посредством DTW на предложенных системах признаков

Для проведения численного исследования был программно реализован алгоритм DTW, при этом в качестве меры близости была выбрана евклидова метрика. В экспериментах участвовало 100 дикторов. Каждый диктор 5 раз произносил каждое слово. В качестве систем признаков использовались признаки (29)-(42). Признаки (31) были построены на основе вейвлета Добеши 4-го порядка, количество уровней разложения P=8, признаки (32) были получены на основе вейвлета Морле при $a_0 = 1.1$; $j_{\min} = 10$; $j_{\max} = 50$.

В табл.1 приведены результаты распознавания речи.

Таблица 1. Результаты численного исследования систем признаков

Система признаков	Вероятность распознавания
коэффициенты КЛП	0.58
коэффициенты отражения КЛП	0.96
автокорреляция КЛП	0.6
кепстр КЛП	0.59
площади поперечных сечений акустической трубы КЛП	0.81
нормированная автокорреляция	0.96
нормированный энергетический спектр КЛП	0.38
меры контрастности КЛП	0.73
нормированный энергетический спектр ДПФ	0.87
меры контрастности ДПФ	0.6
меры контрастности ДВП	0.95
меры контрастности НВП	0.84
нормированное количество импульсов равной длины	0.96
MFCC	0.97

Численное исследование позволяет сделать вывод, что из исследуемых систем признаков при распознавании речи наиболее перспективным являются MFCC, коэффициенты отражения КЛП, нормированная автокорреляция, нормированное количество импульсов равной длины. Однако ложное распознавание составляет не менее 0.03, что для ряда задач может оказаться неприемлемым. Для уменьшения вероятности ложного распознавания предлагается совместное использование систем признаков MFCC и коэффициентов отражения КЛП. В этом случае вероятность ложного распознавания составляет 0.01.

Выводы

Новизна. В статье было проведено численное исследование систем признаков, базирующихся на дискретном преобразовании Фурье; непрерывном и дискретном вейвлет-преобразовании; кепстральном анализе, линейном предсказании; нормированном количестве импульсов равной длины, при этом в качестве метода распознавания был выбран алгоритм DTW. В результате исследования для методики распознавания речи в качестве эффективной системы признаков было выбрано сочетание коэффициентов отражения и MFCC.

Практическое значение. Основные положения работы были использованы при разработке системы распознавания речи, которая может использоваться в ЕЯ-интерфейсе интеллектуальных систем.

Литература

1. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. - М.: Радио и связь,1981. - 496 с.

2. Rabiner L.R., Jang B.H. Fundamentals of speech recognition. – New Jersey: Prentice Hall PTR, Englewood Cliffs, 1993. – P. 507.

3. Редди Д.Р. Машинное распознавание речи // ТИИЭР. – 1976. – Т. 64, № 4. – С. 95-127.

4. Jain A.K., Duin R.P.W., Mao J. Statistical pattern recognition: a review // IEEE Trans. Pattern Anal. Machine Intell. – 2000. – Vol. 22. – P. 4-37.

5. Добеши И. Десять лекций по вейвлетам. – М.: РХД, 2004. – 464 с.

6. Малла С. Вэйвлеты в обработке сигналов. – М.: Мир, 2005. – 671 с.

7. Дорохин О.А., Засыпкин А.В., Червин Н.А., Шелепов В.Ю. О некоторых подходах к проблеме компьютерного распознавания устной речи // Труды Междунар. конф. "Знание-Диалог-Решение" (KDS 97). - Т.1. -Ялта. -1997. - С.234-240.

8. Винцюк Т. К. Анализ, распознавание и интерпретация речевых сигналов. - К.: Наук. думка, 1987. - 261 с.