УДК 004.934.1'1

Е.Е. ФЕДОРОВ

Донецкий государственный институт искусственного интеллекта, Украина

МЕТОДИКА ИДЕНТИФИКАЦИИ ДИКТОРА

Для разработки естественно-языкового интерфейса автоматизированной системы управления (АСУ) в статье предлагается методика идентификации диктора, используемая для выбора словаря эталонов текущего диктора. Для выбора эффективной системы признаков было проведено численное исследование.

методика идентификации диктора, АСУ, естественно-языковый интерфейс, система признаков, дискретное преобразование Фурье, метод кодирования с линейным предсказанием, непрерывное и дискретное вейвлет-преобразование

Введение

Постановка проблемы. В настоящее время актуальной является разработка систем, предназначенных для идентификации диктора. Эти системы имеют широкую область применения – фоноскопическая экспертиза, криптография, охранные системы и др. Применительно к ЕЯ-интерфейсу АСУ они могут использоваться для выбора словаря эталонов текущего диктора. При разработке таких систем важную роль играет выбор системы признаков.

Анализ исследований. В работах [1 – 3] приведены системы идентификации, дающие в большинстве случаев вероятность распознавания ниже 90%. В работе [4] анализу подвергалось все слово, а не конкретные звуки, что сужало область применения.

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

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

– дискретном преобразовании Фурье;

– непрерывном и дискретном вейвлет-преобразовании;

– линейном предсказании;

– нормированном количестве импульсов равной длины, и связанный с этими признаками метод идентификации, основанный на алгоритме динамического искажения времени DTW.

Создание систем признаков для идентификации дикторов

В настоящее время существуют несколько подходов для конструирования систем признаков, используемых при идентификации диктора. 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 ΔN:

$$\hat{s}_n(m) = s_n(m)w(m), \qquad (1)$$

где $w(m)$ – оконная функция, равная нулю вне n-го фрейма.

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:

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

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-м 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]:

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

$$c_{im}=2^{1/2}\sum_{n=0}^{N/2^{i-1}-1}c_{i-1,n}h_{n+2m},$$

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:

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

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:

$$d_{ml}=\sum_{n=0}^{N-1}s(n)\psi_{ml}(n)\Delta t, \quad (5)$$

$$j_{\min}\le m\le j_{\max},\ 0\le l\le N-1,$$

где Δt – величина, обратная частоте дискретизации;

$\psi_{ml}(t)=a_0^{-m/2}\psi\left(a_0^{-m}t-b_0l\right)$, $a_0>1, b\neq 0$,

$j_{\max}, j_{\min}$ – максимальный и минимальный уровни разложения.

Другим подходом выделения акустических параметров, основанным на теории образования речи, является метод кодирования с линейным предсказанием (КЛП).

Линейный предсказатель порядка p с коэффициентами a_k определяется как система:

$$s(n)=\sum_{k=1}^{p}a_k s(n-k), \quad (6)$$

имеющая передаточную функцию:

$$K(z)=\frac{G}{1-\sum_{k=1}^{p}a_k z^{-k}}, \quad (7)$$

где G – коэффициент усиления модели.

Пусть $R_n(i)$ – автокорреляционная функция:

$$R_n(i)=\sum_{m=0}^{N-1-i}s_n(m)s_n(m+i),\ 1\le i\le p, \quad (8)$$

где n – номер фрейма речевого сигнала, i – порядок линейного предсказателя.

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

$$E_n^{(0)} := R_n(0); \quad (9)$$

$$k_i := \left[R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n(i-j) \right] \Big/ E_n^{(i-1)}, 1 \le i \le p; \quad (10)$$

$$\alpha_i^{(i)} := k_i; \quad (11)$$

$$\alpha_j^{(i)} := \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, 1 \le j \le i-1; \quad (12)$$

$$E_n^{(i)} := (1 - k_i^2) E_n^{(i-1)}; \quad (13)$$

$$a_j := \alpha_j^{(p)}, 1 \le j \le p, \quad (14)$$

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 (15), при этом a_0=1.

$$r(0) = \sum_{j=0}^{p} a_j^2, \; r(n) = 2 \sum_{j=0}^{p-n} a_j a_{j+n}. \quad (15)$$

Энергетический спектр определяется как:

$$W(k) = \frac{R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k)}{r(0) - \sum_{n=1}^{p} r(n) \cos\left(\frac{2\pi}{N} nk\right)}, \quad (16)$$

$$k \in \overline{0, N/2 - 1}.$$

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

Другим представлением сигнала является кепстр импульсной характеристики системы линейного предсказания (7). Комплексный кепстр $\hat{h}(n)$ импульсной характеристики получается с помощью рекурсивных соотношений:

$$\hat{h}(n) = a_n + \sum_{k=1}^{n-1} \frac{k}{n} \cdot \hat{h}(k) a_{n-k}, \; n \in \overline{1, p}, \quad (17)$$

где $\hat{h}(n) = a_0$.

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

$$v(m) = \frac{s(m-1) + s(m) + s(m+1)}{3}, \quad (18)$$

после чего вычисляется разность исходного и сглаженного сигналов:

$$\tilde{v}(m) = s(m) - v(m). \quad (19)$$

Сигнал $\tilde{v}(m)$ разбивается на фреймы длиной ΔN. Для каждого n-го фрейма вычисляется d_{nz} – количество импульсов длины z. Вычисления производятся в цикле по $i \in \overline{0, M}$ следующим образом. Полагая изначально $y_n^{(0)}(m) = \tilde{v}_n(m)$, $d_{nz} = 0$, $z \in \overline{1, len}$, согласно (20) определяется длина импульса z, являющаяся числом отсчетов сигнала $y_n^{(i)}(m)$ между двумя соседними локальными максимумами:

$$\begin{aligned} &((y_n^{(i)}(j-1) \le y_n^{(i)}(j) \ge y_n^{(i)}(j+1)) \wedge \\ &(y_n^{(i)}(m-1) \le y_n^{(i)}(m) \ge y_n^{(i)}(m+1)) \wedge \\ &\neg(y_n^{(i)}(k-1) \le y_n^{(i)}(k) \ge y_n^{(i)}(k+1)) \wedge \\ &j < k < m) \to z = m - j. \end{aligned} \quad (20)$$

После чего наращивается счетчик d_{nz} и сигнал $y_n^{(i)}(m)$ подвергается двукратному сглаживанию фильтром:

$$y_n^{(i+1)}(m) = \frac{y_n^{(i)}(m-1) + y_n^{(i)}(m) + y_n^{(i)}(m+1)}{3}. \quad (21)$$

По завершению цикла вычисляется нормированное количество импульсов равной длины согласно следующей формуле:

$$\| d_{nz} \| = d_{nz} \Big/ \sum_{s=1}^{len} d_{ns} \ . \qquad (22)$$

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):

$$X_k = \frac{S^2(k)}{\sum_{i=0}^{N/2-1} S^2(i)}, \ 0 \le k \le N/2-1 . \qquad (23)$$

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

$$X_k = \lg\left(\frac{E_{FFT}(k)}{\sum_{i=0}^{k} E_{FFT}(i)} \right), \ 1 \le k \le L , \qquad (24)$$

где L – количество спектральных полос;

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

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

$$X_k = \lg\left(\frac{E_{DWT}(k+1)}{\sum_{j=1}^{k+1} E_{DWT}(j)} \right), \ 1 \le k \le P \qquad (25)$$

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

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

$$X_k = \lg\left(\frac{E_{CWT}(k + j_{\min})}{\sum_{j=j_{\min}}^{k+j_{\min}} E_{CWT}(j)} \right), \qquad (26)$$

$$1 \le k \le j_{\max} - j_{\min} ,$$

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

5. Коэффициенты линейного предсказания, вычисляемые с помощью алгоритма Дарбина:

$$X_k = a_k , k \in \overline{0, p-1} . \qquad (27)$$

6. Коэффициенты отражения КЛП, определяемые по алгоритму Дарбина:

$$X_i = a_i , \ i \in \overline{1, p} . \qquad (28)$$

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

$$X_k = \hat{h}(k) , \ k \in \overline{1, p} . \qquad (29)$$

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

$$X_{i-1} = A_i \ \ i \in \overline{2, p+1} , \qquad (30)$$

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

9. Нормированная автокорреляция КЛП:

$$X_k = \frac{r(k)}{r(0)}, \ k \in \overline{1, p} , \qquad (31)$$

где $r(k)$ – автокорреляция КЛП, получаемая из (15).

10. Нормированный энергетический спектр КЛП:

$$X_k = \frac{W^2(k)}{\sum_{i=0}^{N/2-1} W^2(i)}, \ 0 \le k \le N/2-1 , \qquad (32)$$

где $W^2(k)$ – энергетический спектр, определяемый из (16).

11. Меры контрастности энергетического спектра КЛП:

$$X_k = \lg\left(\frac{W^2(k)}{\sum_{i=0}^{k} W^2(i)}\right),\ 1 \le k \le N/2 - 1, \qquad (33)$$

где $W^2(k)$ – энергетический спектр, вычисляемый по (16).

12. Нормированная автокорреляция:

$$X_k = \frac{R(k)}{R(0)},\ k \in \overline{1, p}, \qquad (34)$$

где $R(k)$ – автокорреляция, вычисляемая по (8).

13. НКИРД:

$$X_k = \| d_{nk} \|, \qquad (35)$$

где $\| d_{nk} \|$ вычисляются согласно (22).

Метод идентификации диктора

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.

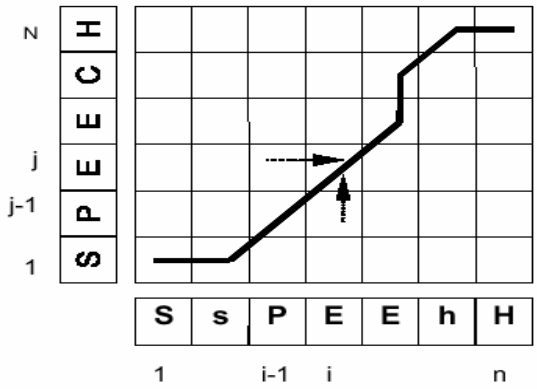

Рис. 1. Сопоставление эталона "SPEECH" и распознаваемого сигнала "SsPEEhH"

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.

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.

Численное исследование по идентификации дикторов 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 ключевое слово. В качестве систем признаков использовались признаки (23) – (35). Признаки (25) были построены на основе вейвлета Добеши 4-го порядка, количество уровней разложения P=8, признаки (26) были получены на основе вейвлета Морле при a_0=1,1; $j_{\min}=10$; $j_{\max}=50$.

В табл. 1 приведены результаты идентификации диктора по ключевому слову |Саша|, фонемам |ш| и |а|.

Таблица 1

Результаты численного исследования систем признаков, используемых ри идентификации диктора

Фонема / Слово	Система признаков	Вероятность идентиф. диктора
1	2	3
слово «Саша»	коэффициенты КЛП	0,84
	коэф. отражения КЛП	0,98
	нормированная автокорреляция КЛП	0,86
	кепстр КЛП	0,82
	площади поперечных сечений акуст. трубы КЛП	0,7
	нормированная автокорреляция	0,76
	нормированный энерг. спектр КЛП	0,6
	меры контрастности КЛП	0,6
	нормированный энерг. спектр ДПФ	0,86
	меры контрастности ДПФ	0,94
	меры контрастности ДВП	0,78
	меры контрастности НВП	0,82
	нормированное количество импульсов равной длины	0,92
	совместное использование коэф. отражения КЛП, нормированного количества импульсов равной длины и мере контрастности ДПФ	0,99
фонема |а|	коэффициенты КЛП	0,36
	коэф. отражения КЛП	0,76
	нормированная автокорреляция КЛП	0,36
	кепстр КЛП	0,4
	площади поперечных сечений акуст. трубы КЛП	0,52
	нормированная автокорреляция	0,54
	нормированный энерг. спектр КЛП	0,24
	меры контрастности КЛП	0,2
	нормированный энерг. спектр ДПФ	0,78
	меры контрастности ДПФ	0,66
	меры контрастности ДВП	0,54
	меры контрастности НВП	0,58
	нормированное количество импульсов равной длины	0,58
фонема |ш|	коэффициенты КЛП	0,68
	коэф. отражения КЛП	0,84
	нормированная автокорреляция КЛП	0,68
	кепстр КЛП	0,58
	площади поперечных сечений акуст. трубы КЛП	0,32
	нормированная автокорреляция	0,34
	нормированный энерг. спектр КЛП	0,3
	меры контрастности КЛП	0,22
	нормированный энерг. спектр ДПФ	0,22
	меры контрастности ДПФ	0,4
	меры контрастности ДВП	0,38
	меры контрастности НВП	0,6
	нормированное количество импульсов равной длины	0,46

Численное исследование позволяет сделать вывод, что из исследуемых систем признаков при идентификации диктора по одной фонеме лучшей системой признаков являются коэффициенты отражения. Для тональных фонем (в данном случае |а|) вероятность идентификации составила 0,76. Для глухих щелевых (или смычно щелевых) фонем (в данном случае |ш|) вероятность идентификации – 0,84.

Из исследуемых систем признаков при идентификации диктора по слову (в данном случае «Саша») наиболее перспективным являются коэффициенты отражения КЛП (вероятность идентификации – 0,98), мера контрастности ДПФ (вероятность идентификации – 0,94), НКИРД (вероятность идентификации – 0,92).

Для повышения вероятности идентификации предлагается совместное использование систем признаков. Вероятность идентификации диктора по слову «Саша», основанной совместном использовании коэффициентов отражения, НКИРД и мере контрастности ДПФ составила 0,99.

Выводы

Новизна. В статье было проведено численное исследование систем признаков, базирующихся на дискретном преобразовании Фурье; непрерывном и дискретном вейвлет-преобразовании; линейном предсказании; нормированном количестве импульсов равной длины, при этом в качестве метода идентификации был выбран алгоритм DTW. В результате исследования для методики идентификации диктора в качестве эффективной системы признаков было выбрано сочетание коэффициентов отражения, нормированного количества импульсов равной длины и мере контрастности ДПФ.

Практическое значение. Основные положения работы были использованы при разработке системы идентификации диктора, которая может использоваться в ЕЯ-интерфейсе АСУ для выбора словаря эталонов текущего диктора.

Литература

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

2. Атал Б.С. Автоматическое опознавание дикторов по голосам // ТИИЭР. – 1976. – Т. 64, № 4. – С. 48-66.

3. Galoonov V.I., Gramnitski S.N., Romashov N.A. VQ and GMM combination for text-independent speaker recognition on telephone channel // SPECOM'2002. – P. 57-60.

4. Алексеев А.С., Федоров Е.Е. Численное исследование систем признаков и методов идентификации // Искусственный интеллект. – 2005. – №3. – С. 557-562.

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.

Поступила в редакцию 16.10.2006

Рецензент: д-р техн. наук, проф. Н.И. Чичикало, Донецкий национальный технический университет, Донецк.