ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ
1971 г. вып. 44

УДК 621.391:681.3.06.

РАСПОЗНАВАНИЕ ФОНЕМ С ИСПОЛЬЗОВАНИЕМ АПРИОРНОЙ ИНФОРМАЦИИ

В.М. Величко

В данной статье описываются эксперименты по распознаванию фонем в отдельно взятых словах. Словарь был подобран так, чтобы все фонемы русского языка встречались в возможно большем числе различных звукосочетаний (этот же словарь использовался в [1]).

Для описания речевого сигнала была принята система признаков, аналогичная использованной в предыдущих экспериментах по распознаванию речи [1,2]. Акустический сигнал пропускался через систему из 4 цифровых октавных фильтров с центральными частотами 900, 1800, 3600 и 7200гц. Параметры применяемых фильтров были выбраны в работе [3]. Слово разбивалось на сегменты фиксированной длительности 14 мсек. Длительность сегмента имеет тот же порядок, что и максимально ожидаемый период основного тона. В каждом сегменте подсчитывалась энергия сигнала в общей полосе E_0 и на выходе каждого из 4 фильтров E_i $(i=1,\dots,4)$. В качестве параметров, характеризующих сегмент, брались величины $x_0 = \ln E_0$, $x_i = \ln \frac{E_0}{E_i}$, т.е. сегмент описывался точкой $\vec{X}$ в 5-мерном пространстве. Слово описывалось как последовательность сегментов $\vec{X}^{(k)}$.

Сравнение сегмента $\vec{X}^{(k)}$ с сегментом $\vec{X}^{(e)}$ проводилось с помощью меры сходства

$$a = \frac{\alpha^2}{\alpha^2 + \rho^2_{ke}},$$

где $\alpha^2 = 2$

$$\rho^2 = \rho_0^2 + \sum_{i=1}^{4} (x_i^{(e)} - x_i^{(k)})^2$$

$$\rho_0 = \begin{cases} |x_0^{(e)} - x_0^{(k)}| - 1 & \text{при } |x_0^{(e)} - x_0^{(k)}| > 1 \\ 0 & \text{при } |x_0^{(e)} - x_0^{(k)}| < 1 \end{cases}$$

Целью работы являлось получение фонетической транскрипции распознаваемого слова. Для достижения этой цели задача распознавания была сформулирована как известная задача об оптимальном размещении точек на отрезке (см., например, [4,5], там же приведена библиография). В работе [6] сведение к задаче размещения используется для членения речевого сигнала на квазистационарные участки.

В данной работе результатом явилось распознавание фонем без предварительного членения распознаваемого слова на фонемы. Принципиальная возможность обойти трудный вопрос о членении на фонемы показана в работах [1,7].

Для выделения эталонов фонем параметры всех слов обучающей последовательности были напечатаны в виде графиков. Затем вручную были выбраны участки слов, соответствующие отдельным фонемам в различных звукосочетаниях. Эталоном фонемы считалась выбранная последовательность сегментов $\vec{X}^{(k)}$. Всего было выбрано 259 эталонов фонем, на каждую из 50 фонем русского языка приходилось от 1 до 8 эталонов из различных звукосочетаний.

Оптимизируемый функционал учитывает, во-первых, величину меры сходства между участками распознаваемого слова и эталонами соответствующих этим участкам фонем и, во-вторых, вероятность сочетаний соседних фонем.

Мера сходства A_n между участком слова и эталоном фонемы определялась как ненормированная длина максимального пути на матрице сходства $\{a_{ke}\}$, элементами которой являются меры сходства a_{ke} между сегментами участка слова $\vec{X}^{(k)}$ и сегментами эталона фонемы $\vec{X}^{(e)}$. Для нахождения длины максимального пути используется метод динамического программирования [2,8]. Изменение выбранных для описания сигнала параметров может быть весьма значительным на длительности фонемы и зависит как от самой фонемы (например, для взрывных, включающих резко отличающиеся части - смычку, взрыв и придыхание), так и от звукосочетания, в которое входит фонема (например, для гласных). Поэтому существенным в принятом определении меры сходства A_n являет-

ся использование представления о "траектории" фонемы, т.е. учет изменения параметров фонемы во времени [2, 9, 10].

Для улучшения качества распознавания учитывается лингвистическая информация о частоте встречаемости сочетаний двух фонем (диад) [11]. В настоящее время получены достаточно полные и надежные данные о двухфонемных сочетаниях русской речи [12]. Однако в данной работе была применена статистика конкретного словаря, которая ввиду специфических требований, предъявленных к словарю, значительно отличается от обычной.

При распознавании слова сначала производилось предварительное определение возможных границ фонем. Общее число границ было взято приблизительно в 3 раза меньше, чем число сегментов в слове, но значительно превышало число фонем. Границы ставились в точках с максимальными $\rho^1_{i-1,i}$ между соседними сегментами, причем каждый интервал между ближайшими границами содержал не менее двух сегментов.

Затем проводилось построение матрицы сходства $f_m(i,\ell)$. Каждый участок слова, начинающийся с i-го интервала и имеющий длину ℓ интервалов, сравнивался с эталонами фонем по сформулированному выше критерию сходства для определения меры сходства $A_{n\,m}$. Все $A_{n\,m}$ упорядочивались по величине, и 4 различные фонемы, имеющие максимальные меры сходства с рассматриваемым участком слова, составляли элементы "матрицы сходства" $f_m(i,\ell)$, $m=1,\dots,4$. Для сокращения довольно громоздкой процедуры вычисления $f_m(i,\ell)$ применялся предварительный отбор эталонов фонем, у которых средние значения параметров на длине эталона близки к соответствующим средним величинам на участке слова. Таким образом, каждый участок слова мог быть распознан как одна из 4-х фонем, на которые он больше всего похож. Оптимальный подбор сочетаний участков производился в соответствии с функционалом, задаваемым рекуррентно с помощью функционального уравнения:

$$F = \max_n [\ln \rho(\neq, n) + F_n(1, N)] \qquad (I)$$

$$F_m(i,k) = \max_{\ell,n} [f_m(i,\ell) + F_n(i+\ell, k-1) + \ln \rho(m,n)].$$

$$1 \le \ell \le \max\{1, \min[6, \ell_{i\,max}(24), M-i+1]\}$$

$$N-k+1 \le i \le M-k+1$$

$$1 \le m, n \le 4$$

$$1 \le k \le N.$$

Начальные условия: $F_n(M+1, 0) = 0$,

при этом $\ln \rho(m,n) = \ln \rho(m,\neq)$.

Здесь

$F_m(i,k)$ – значение функционала при оптимальном членении на k фонем участка слова от i-го интервала до конца слова,

i – номер начального интервала участка слова,

k – число фонем, на которое делится участок слова,

ℓ – число интервалов в участке слова,

M – общее число интервалов в слове,

N – число фонем в слове $(N < M)$,

$\ell_{i\,max}(24)$ – максимальное число интервалов, начиная с i-го, с общим количеством сегментов, не превышающим 24. $\ell_{i\,max}(24)$ вводит ограничение на максимальную длину фонемы ввиду того, что длительность фонемы больше, чем 24x14=336 мсек, маловероятна.

$\ln \rho(m,n)$ – логарифм вероятности следования фонемы n за фонемой m.

$\neq$ означает начало или конец слова.

Функционал (I) учитывает как сходство участков слова с эталонами фонем, так и вероятности сочетания фонем. Величина функционала для оптимального распознавания была найдена методом динамического программирования с использованием "матрицы сходства" $\{f_m(i,\ell)\}$ и таблицы сочетаний диад путем последовательного увеличения k от I до N и перебора по n, ℓ, m, i. Затем восстанавливались оптимальное размещение границ фонем и соответствующие этому размещению фонемы. Число фонем N в слове задавалось переменным, и для каждого N определялся результат распознавания.

Эксперимент по проверке предложенного алгоритма распознавания проводился на универсальной ЭВМ БЭСМ-6. Две последовательности слов, произнесенные одним диктором (мужчина), записывались на магнитофон в тихой комнате с обычной акустикой. Обе последовательности с помощью 9-разрядного аналого-цифрового преобразователя с частотой квантования 20 кгц были введены в ЭВМ

БЭСМ-6 и записаны на магнитную ленту. При этом автоматически определялись границы слова. Затем производилась цифровая фильтрация, выделение логарифмических параметров и запись их на магнитную ленту БЭСМ-6. После этого первая последовательность выводилась на графики для ручного выделения эталонов фонем. Вторая последовательность предъявлялась на распознавание в качестве контрольной. После распознавания печатались варианты фонетической транскрипции слова при разном числе фонем N. Приведем пример печати для слова "лопасть" (диктор произнес это слово как "ЛОП∧С'Т'"):

- ЛОП∧-
- ЛОПАС'Т'
- - ЛОПАС'Т'
- - - ЛОПАС'Т'
- - ЛОПАС'Т'-
- - - ЛЫАЛАС'Т'-

Отобранные вручную лучшие варианты распознавания для каждого слова приведены в таблице I. В таблице указаны предъявленные слова, фонетическая транскрипция лучшего варианта распознавания (смычка опущена), число правильно распознанных фонем в слове $N_г$, число недостающих фонем $N_н$ и число лишних фонем $N_д$. При записи некоторых слов в память БЭСМ-6 отсутствовало начало слова (например, в слове "взрыв"). Это не считалось ошибкой распознавания. Кроме того, не считалась ошибкой замена ударной гласной на безударную и наоборот. Общая надежность распознаваний подсчитывалась по формуле

$$G = \frac{N_г}{N_г + 0{,}5(N_н + N_д)}.$$

При ручном выборе наилучшего варианта распознавания надежность получилась равной 83,6% на материале в 374 фонемы. Работа по автоматическому выбору наилучшего варианта пока не закончена, но было подсчитано число ошибок при выборе варианта, равноудаленного от начала и конца напечатанного списка. (В приведенном примере это четвертое сверху слово "--- ЛОПАС'Т'"). При этом надежность составила 77,5%. Среднее время распознавания одного слова ~ 5 сек.

Для сравнения автоматического распознавания фонем с человеческим распознаванием при одинаковой априорной информации был проведен следующий эксперимент.* На основе статистики диад для русской речи [I2] с помощью таблицы случайных чисел был составлен словарь из 47 бессмысленных слов. Словарь был записан на магнитофон в тех же условиях и тем же диктором, что и при автоматическом распознавании, и предъявлен 5 аудиторам. Разрешалось многократное проигрывание записи по просьбе аудиторов. Надежность распознавания, подсчитанная по формуле (4), колебалась для разных аудиторов от 86,3% до 95% и в среднем составила 90,I%. Нижний порог человеческого распознавания в этом эксперименте довольно близок к полученной надежности автомата, но следует отметить, во-первых, использование различной статистики в этих экспериментах и, во-вторых, непроизвольное использование человеком дополнительной априорной информации, что затрудняет сравнение результатов.

В заключение автор выражает благодарность Н.Г.Загоруйко за руководство работой и полезные советы, Л.С.Юдиной за составление и транскрибирование словаря, Э.Х.Гимади и В.Д.Гусеву за обсуждение алгоритма.

*) Идея эксперимента и методика его проведения предложены Н.Г. Загоруйко.

Таблица I

Слово	Транскрипция	$N_п$	$N_и$	$N_э$
помпа	помпʌ	5	0	0
свист	св'ист	5	0	0
лопасть	лопʌс'т'	6	0	0
жемчужина	жамчужын	8	I	0
фельдфебель	ф'ьл'тф'эб'ьл'	9	0	0
дифференциация	г'иф'эр'энцыацыа	I2	2	I
смех	см'эх	4	0	0
специфика	п'ицып'икʌ	7	I	I
устройство	уростʌ	6	4	0
взрыв	грыф	3	I	I
гегемон	гиjэмон	6	I	I
груздь	грус'т'	5	0	0
гимн	г'имн	4	0	0
кизил	п'из'ил	4	I	I
медянка	гн'ид'анкʌ	6	I	2
лемех	л'эм'эх	5	0	0
лыжи	лыжы	4	0	0
двухвостка	двопопʌ	4	6	3
скипетр	к'ип'ътр	6	0	0
хищник	х'ищ'н'ик	6	0	0
жилы	жэ	I	3	I
чешуя	чашуjа	6	0	0
шихта	шухтʌ	4	I	I
чечевица	чичэда	5	3	I
резерв	к'из'эф	4	2	I
бегемотики	б'иг'имот'ик'э	9	I	I
якив	п'эф	I	2	2
ёлзик	ф'из'ик	5	0	0

Продолжение таблицы I

Слово	Транскрипция	$N_п$	$N_и$	$N_э$
щавель	щ'ав'э	4	I	0
щебень	щ'эб'имн'	5	0	I
ходьба	ф'эд'ба	3	2	2
щуплый	щ'уплыj	6	0	0
буддизм	буд'изм	6	0	0
жужелица	жужэцʌ	6	2	0
бубен	бубэн	5	0	0
дуги	дуг'ър'	4	0	I
бурнус	бус	3	3	0
безденежье	б'ьад'эн'ьжjэ	I0	0	0
кимограф	х'имогрʌ	6	2	I
гнездовье	гн'издон	6	3	I
хвощ	лощ'	2	I	I
хилость	х'илʌс'т'	6	0	0
пикет	п'ик'э	4	I	0
химик	х'им'ък	5	0	0
питомник	п'иг'ъмн'ък	6	2	2
грязища	б'ир'эа'ищ'ʌ	6	I	2
пигмей	п'игм'ик	4	2	2
сдоба	здопʌ	4	I	I
пичуга	п'ичупʌ	5	I	I
певица	п'ив'ицʌ	6	0	0
воздух	лоздух	5	I	I
дверь	д'э	I	3	I
гнев	гн'эм'э	3	I	2
Щецин	щ'эцын	5	0	0
бирюза	б'ьр'ьжна	4	2	3

Продолжение таблицы I

Слово	Транскрипция	$N_п$	$N_м$	$N_л$
ивал	вал	3	I	0
купальца	к'уплаба	5	2	I
кефаль	к'ьфа	4	I	0
гистохимия	дв'исто/эм'ик^	7	3	5
хихиканье	х'их'ыкалин	5	4	4
		304	70	49

Литература

1. Загоруйко Н.Г., Величко В.М., Волошин Г.Я., Гусев В.Д., Ёлкина В.Н., Бахмутова И.В., Хайретдинова А.Г., Юдина Л.С. Эксперименты по автоматическому распознаванию речевых сигналов. Труды АРСО-IУ. Киев-Канев, 1968.
2. Величко В.М., Загоруйко Н.Г. Автоматическое распознавание ограниченного набора устных команд."Вычислительные системы", Новосибирск, вып. 56, I969.
3. Курилов Б.М., Гаврилко Б.П. Сжатие описания сигнала и членение речи на фонемы. "Вычислительные системы", Новосибирск, вып. 56, I969.
4. Дементьев В.Т. Об одной задаче оптимального размещения точек на отрезке. "Дискретный анализ", Новосибирск,вып.4, I965.
5. Гимади (Гимадутдинов) Э.Х. Выбор оптимальных шкал в одном классе задач типа размещения, унификации и стандартизации. "Управляемые системы", Новосибирск, вып. 6, I970.
6. Винцюк Т.К. Оптимальное разбиение последовательности элементов на подпоследовательности. "Кибернетика", Киев,вып.4, I970.
7. Загоруйко Н.Г. Алгоритм распознавания фонем по последовательности сегментных решений. Труды АРСО-IУ,Киев-Канев, I968.
8. Беллман Р., Калаба Р. Динамическое программирование и современная теория управления. "Наука", Москва, I969.
9. Бондарко Л.В., Загоруйко Н.Г., Кожевников В.А., Молчанов А.П., Чистович Л.А. Модель восприятия речи человеком. "Наука" Новосибирск, I968.
I0. Голубцов С.В. Труды АРСО-IУ. Киев-Канев, I968.
II. Волошин Г.Я. Об использовании языковой избыточности для повышения надежности автоматического распознавания речевых сигналов. "Вычислительные системы", Новосибирск , вып. 28, I967.
I2. Ёлкина В.Н., Юдина Л.С., Хайретдинова А.Г. Статистика двух- и трехфонемных сочетаний русской речи. "Вычислительные системы", Новосибирск, вып. 37, I969.

Поступила в редакцию
20.I2.1970 г.