ЭКСПЕРТНЫЕ СИСТЕМЫ И РАСПОЗНАВАНИЕ ОБРАЗОВ
(Вычислительные системы)
1988 год Выпуск 126

УДК 621.391

# ЭКСПЕРТНАЯ СИСТЕМА, ИСПОЛЬЗУЮЩАЯ ЗНАНИЯ О МОРФЕМАХ, ДЛЯ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ СЛИТНО ПРОИЗНОСИМЫХ ЧИСЕЛ (ОТ 0 ДО 999)

З.А.Вадова, А.И.Зеленый,
Р.В.Саруханян, В.Н.Трунин-Донской

Одним из методов, используемых при декодировании (автоматическом распознавании) слитной речи, является метод, формализующий опыт чтения спектрограмм экспертами по фонетике.Этот метод привлекает к себе все возрастающий интерес. Так, формализацией правил фонетического декодирования в СССР занимаются в МГУ [1], ВЦ АН СССР [2] и других отечественных организациях, а также за рубежом: в США [3,4], Канаде [5,6], Франции [7,9],Японии [9], Великобритании [10]. Обычно искусственную систему,моделирующую на ЭВМ деятельность специалиста-эксперта в какой-либо предметной области, называют экспертной системой.

В ВЦ АН СССР разрабатывается система для многодикторного автоматического распознавания слитно произносимых чисел. Эта система использует знания экспертов-лингвистов, расшифровывающих вручную распечатки матриц признаков, соответствующих произнесенным высказываниям. Параметрами первичного описания являются признаки, получаемые при помощи устройства выделения речевых признаков УВРП-М, подробно описанного в [11]. Наряду с гребенкой из 15 аналоговых частотных полосовых фильтров используются каналы, выделяющие формантные параметры, общую энергию

сигнала, энергию в низкочастотной и высокочастотных областях, а также ряд параметров клиппированой речевой волны.

Аппаратурно-программная система, реализованная на измерительно-вычислительном комплексе ИВК-2, позволяет эксперту получить гистограммы, отдельных параметров звуков и звукосочетаний, а также отрабатывать, используя отображаемые на дисплее и листингах признаки, логико-лингвистическую стратегию принятия решения при декодировании слитных словосочетаний. Правила де - кодирования матриц, соответствующих словосочетаниям (динамических спектрограмм, к которым добавляются дополнительные параметры, поступающие с УВПР-М), используются в качестве основных логико-лингвистических правил принятия решения.

Разработаны алгоритмы квазифонетической сегментации сло - восочетаний. Наличие в словаре ограниченного набора морфем позволили перейти от метода, основанного на использовании фонем и звукосочетаний, к морфемному методу, использующему дополни - тельные интегральные логические признаки. В основе фреймов,последовательностью которых перекрывается высказывание, лежат последовательности участков речи, составляющих морфему, причем эти участки могут быть либо фонемой, либо ее частью,либо несколькими фонемами в зависимости от работы алгоритма сегментации.

Основных морфем, составляющих рабочий словарь, 18: но́ль, оди́н, два́ (две́,две), три́ (три), четы́ре (четы́р, чты́р), пя́ть, (пят), ше́сть (шес, шизь, шесь), се́мь (сем), во́семь (восем), де́вять (девят, девя), де́сять, на́дцать (а́дцать, а́цать), цать, со́рок, ся́т (сят), но́ста, сто́ (ста, сти), со́т. В скобках приведены варианты морфем. Ударные гласные помечены знаком ударения. В связи с большой изменчивостью первичных параметров на морфемах рабочего словаря используются различные базовые формы ударных и безударных морфем, причем характер базовой формы определяется не только положением морфем в изолированном слове, но и ее положением относительно ударного слога (смыслового уда-

рения) всего словосочетания.В двух-, трехсловных слитных сочетаниях лексическое (морфемное) ударение не всегда соответствует смысловому. Как правило, большинство дикторов делают смысловое ударение на третьем слове в трехсловных сочетаниях (или на втором в двухсловных). Наименее четко обычно произносится второе слово трехсловных сочетаний. Первое слово чаще всего артикулируется достаточно четко, хотя степень ударности гласного выражена слабее, чем на завершающем слове.

Для обнаружения и верификации морфем неизвестного словосочетания используются следующие примеры:

а) грубая сегментация 1-го уровня и разбиение словосо - четаний на участки "тональный - шумный - пауза";

б) более точная классификация отрезков речевой волны - шумных и тональных;

в) определение опорных звукосочетаний типа "ац", "иц", "ст", "сть", выделяемых с высокой точностью по их параметрам;

г) поиск морфем на отрезках речи, включающих опорные звукосочетания; при этом используются знания о статистике и динамике параметров первичного описания на отрезках, составляющих морфему, а также знания о ее длительности и общих характеристиках;

д) ранжирование вероятностей возможных морфем в соответ - ствии с общей длительностью словосочетания и графом порождения эталонов трехзначных чисел, представленных на рис.1.

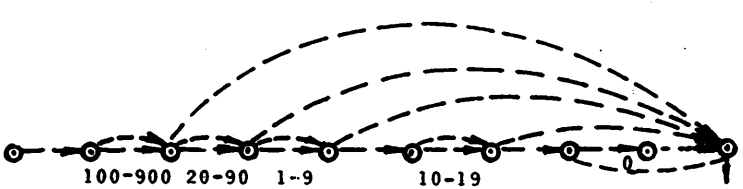


Рис. 1

Вообще говоря, наш подход можно было бы назвать фонемно-морфемным, так как предварительная сегментация слитной речи осуществляется на фонемном (точнее, квазифонемном) уровне. Но мы называем его морфемным потому, что в нашем случае эксперт-фонетист легко вырабатывает для них решающие правила. Количество морфем в таких малых проблемно-ориентированных словарях, как наш, ограничено, а поэтому каждая из гипотезируемых морфем может быть легко верифицирована. При поиске островков фонетической надежности в слитной речи морфемы, во-первых, перекрывают бóльшую часть высказывания, чем отдельные звуки, во-вторых, позволяют с бóльшей точностью прогнозировать звукосочетания на соседних участках, и, в-третьих, их наличие подтверждается большим числом признаков, характеризующих составляющие их звуки, а значит, морфемы могут быть более надежно распознаны.

Рассмотрим первый этап автоматического распознавания - квазифонетическую сегментацию и маркировку слитной речи, базирующуюся на моделировании деятельности эксперта-фонетиста. Сегментацию в дальнейшем используют для выявления морфем и последующей дешифровки высказывания, используя граф рис.1. Исходными данными для программы сегментации являются: массив признаков **NØ** (признак мгновенной частоты, усредненной на интервале 10 мс) и массив признаков гласности **SUNG**, который получается программой суммированием чисел, каждое из которых представляет собой амплитуду в одном из 15 полосовых фильтров. В формировании значений **SUNG** используются только первые 10 фильтров гребенки, т.е. выбирается область, примерно соответствующая полосе первой форманты.

После получения речевой реализации непосредственно с микрофона или из архива вышеуказанные массивы обрабатывают путем:

а) сглаживания признаков на окне в 50 мс;

б) определения максимальных значений признаков **NØ** и **SUNG** в данной реализации;

в) нормирования массивов по соответствующему максимуму. Сглаживание и нормирование признаков производится с целью устранения неинформативных изменений и уменьшения влияния громкости произношения;

г) формирования специального массива OBRAZ , в котором буквой S обозначается позиция (номер отсчета),где NØ больше порога, а буквой Г обозначается позиция, в которой признак гласности SUNG превысил пороговое значение. Все остальные позиции массива OBRAZ , т.е. не прошедшие через тот или иной порог, отмечаются символом 0. Пороги для обоих признаков устаналиваются экспериментально.

На этом заканчивается первый уровень сегментации. Например, для одной из реализаций морфемы "четыр" массив OBRAZ имеет вид: SSSSSSS ГГГГГГ ØØØØ SSS ГГГГГГГГГГГГ Ø ГГ. Границы сегментов являются метками первого уровня,и они всегда парные.

На втором уровне сегментации выполняются следующие действия:

а) определяются производные сглаженных массивов признаков NØ и SUNG ;

б) определяются экстремумы производных;

в) фиксируются номера отсчетов, в которых хотя бы один из экстремумов превысил порог. При этом если расстояние между экстремумами менее 3 отсчетов, то фиксируется лишь один из них, т.е. не допускаются сегменты,имеющие длительность менее 30 мс.

Адреса экстремумов в виде номеров отсчетов являются метками второго уровня. Они необходимы для уточнения границ сегментов, определенных на первом уровне сегментации, а также в тех случаях, когда через порог не прошли те или иные тональные или шумные участки высказывания. Такое явление наблюдается при недостаточно громком произношении.

Массивы границ сегментов, определенных на первом и втором уровнях, являются исходными для подпрограммы коррекции. В ее задачу входят объединение границ обоих уровней и коррекция массива OBRAZ . На этом этапе выполняются следующие действия:

а) для каждой пары меток первого уровня (начала и конца сегмента) ищутся метки второго уровня,близкие (в пределах трех отсчетов) к меткам первого уровня,и найденные метки помечаются в массиве границ второго уровня как ненужные;

б) если метка второго уровня лежит вне сегмента,определенного на первом этапе сегментации, то формируется новый сегмент при условии превышения порога значениями NØ и SUNG в окрестности данной метки. В противном случае данная метка также игнорируется;

в) все метки второго уровня, лежащие внутри сегмента и отстоящие от границ более чем на три отсчета,считаются разделяющими.

До сих пор в массиве OBRAZ отмечены как шумные или тональные только те отсчеты, которые превысили пороги. Все остальные отмечены нулем. Следующий шаг - это определение сегментов слабого тона типа L везде, где SUNG больше 4. К сегменту типа L будут с большой вероятностью отнесены все звонкие согласные, а также гласные со слабой энергией. Для выделения "и" введен еще один тип сегмента Y . Это сегмент типа L , но с высокой второй формантой.

Теперь OBRAZ почти полностью сформирован, остается произвести коррекцию конца. Дело в том, что в конце высказывания часто бывает слабый тон от придыхания. Это выражается в виде такой, например, "рваной" последовательности, как LØLLØL... Среди этих сегментов может пройти незамеченным слабый взрыв на "т" или "к". Поэтому алгоритм поиска конца для каждого отсчета с конца,помеченного как L или 0, ищет в массиве NØ взрыв с достаточно низким порогом, и если он найден, то поиск закан-

чивается. Если он не найден, а обнаружен сегмент, отличный от L или 0, то последний L-сегмент восстанавливается (в случае звонкой согласной на конце). Здесь помогает введенный ранее Y-сегмент. Не будь его, могли быть утеряны слабые "и" в конце реализации.

Окончательная сегментация делается следующим образом: начало и конец каждого сегмента заносятся в массив общей сегментации в том случае, если длительность сегмента не менее 30 мс. Таким образом устраняются по мере возможности переходные сегменты. Если в пределах границ сегмента имеются метки второго уровня, то сегмент разделяется. Как показывает практика, это не всегда оправдано, так как могут дробиться ударные гласные сегменты. Можно избежать этого увеличением окна сглаживания и уменьшением порогов, но при этом возникает другая проблема: в один сегмент с гласными попадают сонорные и звонкие шумные. Поэтому приходится идти на компромисс.

При анализе статистического материала экспертами-лингвистами получены гистограммы распределения длительностей одно-, двух- и трехсловных сочетаний чисел (для 10 дикторов мужчин и женщин). Кривые распределения, которые в дальнейшем используются для декодирования слитной речи, приводятся на рис. 2 (1 - однословные; 2 - двухсловные; 3 - трехсловные).

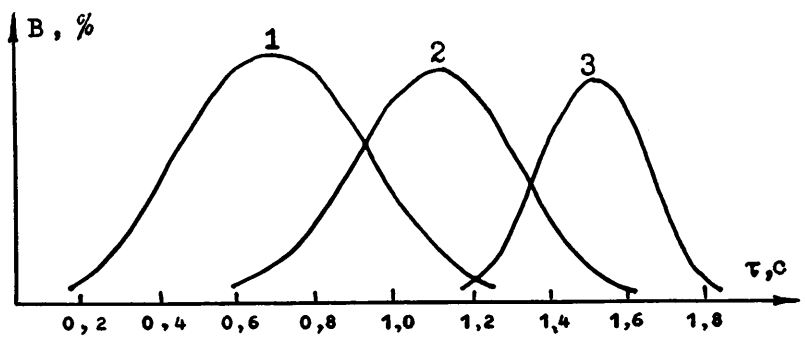


Рис. 2

Следует отметить, что наибольшие трудности для экспертов-лингвистов при декодировании слитной речи представляют участки, содержащие только звонкие звуки, т.е. отрезки речи без щелевых, аффрикат и глухих взрывных. При классификации звонких звуков используется более точный анализ отрезков высказывания, о котором говорилось ранее (сегментация и квазифонетическая маркировка).

Рассмотрим в связи с этим "тяжелый случай" классификации и маркировки отрезков слова "ноль", в котором присутствуют лишь звонкие участки. Параметрическое представление последовательности фонем, составляющих морфемный эталон, имеет вид:

$$\text{ноль} \to V(\tau \geq 350 \text{ мс}, \exists R) \to \left\{ V_1 \left[ \left( \frac{I_{ср}}{I^{морф}_{max}} \leq 0,4; \right. \right. \right.$$

$$\left. \tau \geq 0,8\,T \right) \wedge (F1, F2 < 1000 \text{ Гц}) \wedge (N\emptyset \leq 5; \tau \geq 0,8\,T) \wedge$$

$$\wedge \sum_{i=9}^{15} A_i \leq 5 \text{ у.е.}; \tau \geq 0,8\,T) \Big] > V_2 [(\exists I^{морф}_{max}) \wedge$$

$$\wedge \left( \frac{I_{ср}}{I^{морф}_{max}} \geq 0,7 \right) \wedge \left( \sum_{i=6}^{10} A_i \geq 0,5\,S \right) \wedge (\tau \geq 100 \text{ мс}) \wedge$$

$$\wedge (F1_{max} \geq 600 \text{ Гц}) > V_3 \left[ \left( \frac{I_{ср}}{I^{морф}_{max}} \leq 0\ 5 \right) \wedge (F2 = \right.$$

$$\left. \left. = 1400\text{-}2000 \text{ Гц}) \wedge \left( \sum_{i=3}^{4} A_i > \sum_{i=1,2,5,6,\ldots,15} A_i \right) \right] \right\}.$$

Эта запись означает, что морфема "ноль" - звонкий отрезок речи длительностью более 350 мс, на котором полностью отсутствует шумовая составляющая. Этот отрезок может быть раз-

бит на три звонких участка. Первый участок не является ударным гласным и на 80% своей длины имеет интенсивность, составляющую менее 0,4 от максимальной интенсивности морфемы. Значения первой и второй формантных частот - ниже 1000 Гц, значение усредненной мгновенной частоты - не выше 500 Гц. Более чем на 80% отсчетов этого звонкого участка суммарная энергия в семи высокочастотных каналах гребенки фильтров менее 5 условных единиц. Второй участок, соответствующий ударному "о", характеризуется наличием сегмента с максимальным значением интенсивности морфемы, относительно высокой громкостью всех отсчетов ($I_{ср}/I_{max} < 0,7$), сильной энергетической составляющей в полосе от 400 Гц до 1,25 кГц (более половины всей энергии), относительно высоким максимальным значением частоты первой форманты, типичным для аллофона "о" в сочетании "ноль", и значительной длительностью ($\tau \geq 100$ мс). На третьем участке наблюдается падение интенсивности ($I/I_{max} < 0,5$), повышение до 1400 Гц и выше частоты 2 и высокая общая энергия в полосе нижних частот 200-400 Гц.

Аналогичные представления составлены для всех морфем словаря. Предполагается использовать эти эталонные записи для автоматического распознавания чисел от 0 до 999, произносимых произвольным диктором.

## Литература

1. ЗИНОВЬЕВ Н.В., ЗАХАРОВ Л.М., АМПИЛОВ В.В.Методика чтения "слепых" сонограмм //Автоматическое распознавание слухо - вых образов. Тез. докл. АРСО-12. Т. 3. 1982. -С. 351-353 (ИК АН УССР).

2. ТРУНИН-ДОНСКОЙ В.Н. Экспертные системы и фонетическая маркировка сегментов слитной речи //Анализ, распознавание и синтез речи. - М., 1987. - С. 3-16 (ВЦ АН СССР).

3. JOHANNSEN J. et al. A speech spectrogram expert //Proc. IC ASSP-83. - 1983.- Vol. 2. - P. 746-749.

4. ZUE V., LAMEL L.F. An expert spectrogram reader: A knowledge-based approach to speech recognition //Proc. IC ASSP-86. - 1986. -Vol. 2. - P. 1197-1200.

5. DE MORI R. et al. Integration of acoustic, phonetic, prosodic and lexical knowledge in expert system for speech understanding //Proc. IC ASSP-84. - 1984. -Vol. 3. - P. 42.9.1-42.9.4.

6. DE MORI R., LAM L. Plan refinement in a knowledge-ba - sed system for automatic speech recognition //Proc. IC ASSP-86. - 1986. - Vol. 2. -P. 1217-1220.

7. HATON J.P., DAMESTOY J.P. Afframe language for the control of fonetic decoding in continuous speech recognition //Proc. IC ASSP-85. - 1985. -Vol. 4. - P.1565-1568.

8. GARBONELL K. et al. APHODEX, design and implementation of an acoustic-phonetic decoding expert system //Proc.IC ASSP-86. - 1986. - Vol. 2. - P. 1201-1204.

9. MIZOGUCHI R., TSUJINO K., KAKUSHO O. A continuous speech recognition system based on knowledge engineering tech - niques //Ibid. - P. 1221-1224.

10. GREEN P.D., WOOD A.R. A representational approach to knowledge-based acoustic-phonetic processing in speech recognition //Ibid. - P. 1205-1208.

11. КОЗАДАЕВ В.П., НГУЕН АНЬ ТУАН, РОДИОНОВА Г.Г. ,ТРУНИНДОНСКОЙ В.Н. Выделение новых интегральных признаков способа и места образования звуков //Анализ речевых сигналов. М., 1984. - С. 41-49 (ВЦ АН СССР).

Поступила в ред.-изд.отд.

8 февраля 1988 года