# АВТОМАТИЧЕСКАЯ СЕГМЕНТАЦИЯ СОЧЕТАНИЙ ЗВУКОВ (ДИАД)

ПЁТР ДОМАГАЛА

Лаборатория Акустической Фонетики
Институт Основных Проблем Техники ПАН
Познань

РЕЗЮМЕ

Представлена система сегментации сигнала речи на основании изменений распределений энергии в соседних спектрах. Применён алгоритм, основанный на нескольких логических зависимостях, что требует низких вычислительных мощностей ЭВМ. Система применена для сегментации 444 польских диад, произнесённых 8 дикторами. Получено в среднем 90% правильных сегментаций.

## ВВЕДЕНИЕ

Исследования, касающиеся автоматической сегментации сигнала речи (АСР), обычно представляют собой предварительный и существенный этап сложного процесса автоматического распознавания речи (АРР). В области проблематики анализа речевых сигналов человеком или машиной понятие сегмента применяется, вообще говоря, для определения временно́го фрагмента единицы (обычно лингвистической), являющейся элементом во множестве знаков, подвергающихся распознаванию. Таким множеством может быть слово или фраза. В таком случае сегментом может быть звук, сочетание двух звуков (диада), либо слог. В случае модели распознавания биологической системой сегмент не может выбираться произвольно, а должен соответствовать фактическим физиологическим, неврологическим и психологическим процессам, имеющим место в восприятии. При автоматическом распознавании не обязательно копирование этих процессов, и поэтому сегментом может быть фрагмент сигнала, который не совпадает ни с элементом восприятия, ни с лингвистическим элементом, разве что предполагается, что в рамках проблематики бионики специально конструируются системы, копирующие биофизические функции. Несмотря на то, что в настоящее время существует много моделей и теорий восприятия речи /3/, часто противоречащих одна другой, имеется значительное единство мнений, касающееся того, что в нормальном процессе восприятия речи человеком в рамках таких более крупных лингвистических единиц как предложение (высказывание) или слово, происходит выделение по крайней мере некоторых сегментов, длительностью приближающихся к звуку. В то же время не ясно, на каком уровне восприятия происходит звуковая сегментация - слуховом, фонетическом или высшем. Сегмент является элементом конечного множества значительно меньшей численности, чем численность множества элементов высшего ряда -для данного законченного словаря число слов гораздо больше числа слогов, которое, в свою очередь, больше числа фонем.
В настоящей работе понятие сегмента отождествляется с фонетической единицей, понимаемой как моносегментный звук, либо с сегментом полисегментного звука. Представлена АСР, опирающаяся на анализ изменений распределения энергии в спектре, а также результаты сегментации чаще всего встречающихся в польском языке диад.

## ОПИСАНИЕ МЕТОДА

Для реализации метода была использована аналого-цифровая система, в состав которой входят: 60-канальный анализатор спектра, интерфейс, соединяющий аналоговый источник сигнала с микро-ЭВМ и микро-ЭВМ МЕРА-303. Аналоговый анализатор спектра имеет 43 полосы постоянной ширины, составляющей 80 гц, покрывающие область частот от 120 до 3560гц,

а также 17 полос с шириной, зависящей линейно от среднегеометрической частоты и покрывающей диапазон от 3560 гц до 7000 гц. Выходы отдельных каналов циклически подключаются к общему выходу. Полученная цифровая спектрограмма сигнала речи после усреднения остаётся в памяти ЭВМ в виде таблицы с координатами времени и частоты /1/. Для каждой пары очередных спектров $K-1$ и $K$ создан $N$ - элементный ряд ($N$ - число каналов) с элементами $\gamma_{iK} = a_{iK-1} - a_{iK}$, где $i = 1,\ldots,N$ обозначает номер полосы, $K$ - очередной квант времени. Ряд $\gamma_{iK}$ был поделён на $c(k)$ составных рядов при применении критерия согласованности знака и достаточно высоком абсолютном значении, то есть направления и скорости изменения уровня. Через $z(k)$ обозначен знак элементов последнего составного ряда, принимая 0 для положительных величин и 1 для отрицательных. Принято, что граница между сегментами будет обозначена в следующих случаях:

1) $c(k) = 1 \wedge [c(k+1) \neq 1 \vee [c(k+1) = 1 \wedge z(k) \neq z(k+1)]]$
2) $c(k) = c(k+i) = 1 \wedge z(k) = z(k+i) \wedge [c(k+n) \neq 1 \vee [c(k+n) = 1 \wedge z(k) = z(k+n)]]$ где $i=1, 2, .., n-1$
3) $c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1\ c(k+1) \neq 2$
4) $c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge z(k) \neq z(k+n) \wedge c(k+n) = 2$ где $i=0, 1, \ldots, n-1$ а $n \geqslant 1$
5) $c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge c(k+n) \neq 1 \wedge c(k+n) \neq 2$ где $i=0, 1, \ldots, n-1$ а $n \geqslant 2$.

Вышеуказанные логические зависимости были введены в систему МЕРА 303. Для подготовки экспериментального материала были использованы данные, полученные Яссемом, Лобач в их работе, касающейся фонотактики польского языка /2/. Опубликованный в этой работе список чаще всего встречающихся в польском языке диад в 94% охватывал $5 \cdot 10^5$ анализируемую там выборку численностью $10^5$. Этот список был использован с исключением диад типа: "#F", "$F_H$" и "$F_j F_j$" (# обозначает паузу, $F$ - какую-либо фонему, а "$F_j F_j$" обозначает диаду, состоящую из одинаковых фонем). Это обозначало сокращение списка до 444 пар различающихся между собой фонем.

Для каждой диады было образовано искусственное слово (логатом), содержащее две её реализации и отвечающее принципам фонотактики польского языка. Созданные логатомы были сгруппированы в четыре списка в очерёдности, соответствующей порядку появления диад в списке частотности. Отдельные списки содержали 112, 109, 105 и 118 логатомов каждый. Каждый логатом был произнесён по 3 раза восьмью дикторами (5 мужских голосов и 3 женских голоса) и записан на магнитную ленту. Представленный выше материал был подвержен автоматической сегментации. В качестве порогового значения скорости изменения уровней сигнала в отдельных каналах спектрального анализатора были приняты примерно 30 дБ/23 мсек (23 мсек - это временное расстояние между соседними спектрами). Это значение было постоянным для всех голосов. Была проанализирована частота и проведена автоматическая сегментация каждого воспроизведённого с магнитной ленты логатома. Спустя примерно 1,5 сек на экране монитора появлялась спектрограмма высказывания с обозначенными границами. Изображение оставалось неподвижным в течение примерно 5 секунд, т.е. до времени появления следующей спектрограммы. В это время следовало найти те фрагменты спектрограммы, которые относились к диадам, являющимся основой конструкции логатома. Если положение автоматически определённой границы совпадало с серединой переходного участка между двумя фонемами с точностью расстояния между двумя соседними спектрами (23 мсек), то сегментация считалась правильной. Это расстояние (в два раза большее, чем применяемое в других методах) превышает длительность самых коротких артикуляционных явлений. Условием положительной оценки сегментации диад, содержащих полисегментальную фонему, было выделение основного сегмента, например, смычки в согласных смычных звуках. Каждый логатом произносился по три раза одним и тем же диктором и, следовательно, каждая диада была произнесена 6 раз. Записывались результаты 5 первых повторений.

## ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В Табл. 1 представлены средние значения эффективности АСР по отдельным голосам, для мужских голосов и женских (м,ж), а также для целой группы (мк). Значения, находящиеся в колонках, обозначенных "а", являются средними арифметическими, полученными в эксперименте, а данные в колонках, обозначенных "б", являются средними, учитывающими распределение частоты встречаемости анализированных диад в польском языке. В Табл.2 представлены средние значения и дисперсии отсутствующих сегментаций, приходящихся на одну диаду. Диады, чаще встречающиеся в польском языке (Список 1), легче поддаются сегментации, чем диады, встречающиеся реже (Список 4). Из рис. 1 вытекает, что диады, сегментируемые с эффективностью не менее 80% (в среднем 1 ошибка на голос) составляли 75% всех диад. Принимая во внимание частотность, устанавливаем, что величина эта возрастает на 5%. Для всего материала (444 диады×8 дикторов×5 повторений) получено 90% эффективности сегментации. В Табл. 3 представлено распределение типов диад из анализируемого материала вместе с данными, касающимися процента отсутствующих сегментаций. (С обозначает согласный, $V$ - гласный, $\check{V}$ - не образующие слогов $j, \omega$). Самой большой податливостью на сегментацию отличаются диады типа $CV$ и $\check{V}C$. Причины низкого уровня правильной сегментации ниже 60% для некоторых диад следующие:

- малая контрастность /eo, nu, un, ɲu, um, oa, uω, ʃʃ, oω, ao, ea, ɛ̇j, xf, ji, ωu, ij, fs/;
- низкий энергетический уровень обеих фонем /ωt, nts, vu, nm, ug, ndz, gr, mr, nt, ng, ur, ωt, mp, nd, mb/.

Первая причина имеет объективный характер, вторая вытекает из специфики метода. Представленный и применённый метод АСР характеризуется относительно высоким уровнем эффективности сегментации при использовании минимальных вычислительных мощностей, что гарантирует работу в условиях, приближённых к реальному времени. Метод можно использовать в качестве предварительного этапа процедур параметризации и распознавания речи. Полученные результаты для чаще всего встречающихся в польском языке диад можно использовать для создания искусственных языков для целей коммуникации человек - машина.

| Голос | Список 1 | | Список 2 | | Список 3 | | Список 4 | | Вместе | |
|---|---|---|---|---|---|---|---|---|---|---|
| | а | б | а | б | а | б | а | б | а | б |
| 1м | 90,8 | 90,4 | 90,6 | 90,0 | 88,3 | 89,6 | 89,3 | 88,6 | 89,8 | 90,1 |
| 2м | 94,0 | 95,8 | 92,1 | 92,4 | 87,4 | 86,5 | 84,6 | 84,2 | 89,5 | 93,6 |
| 3м | 83,6 | 82,8 | 88,4 | 88,4 | 80,4 | 79,2 | 81,7 | 82,4 | 83,5 | 83,5 |
| 4м | 96,6 | 96,0 | 94,1 | 94,4 | 90,6 | 90,0 | 93,9 | 94,0 | 93,8 | 95,0 |
| 5м | 96,1 | 96,5 | 81,5 | 82,1 | 89,3 | 89,4 | 80,2 | 79,8 | 86,7 | 91,9 |
| 6ж | 89,3 | 88,6 | 87,2 | 87,2 | 85,7 | 84,1 | 82,8 | 83,5 | 86,2 | 87,7 |
| 7ж | 91,1 | 90,2 | 83,7 | 83,6 | 70,0 | 71,1 | 84,0 | 83,7 | 82,4 | 86,6 |
| 8ж | 92,1 | 92,1 | 87,7 | 87,7 | 80,0 | 79,8 | 77,4 | 78,9 | 84,4 | 89,2 |
| м | 92,2 | 92,3 | 89,3 | 89,5 | 87,2 | 86,9 | 85,9 | 85,8 | 88,6 | 90,8 |
| ж | 90,8 | 90,3 | 86,3 | 86,2 | 78,6 | 78,4 | 81,4 | 82,0 | 84,3 | 87,8 |
| мк | 91,7 | 91,5 | 88,2 | 88,2 | 84,0 | 83,7 | 84,2 | 84,4 | 87,0 | 89,7 |

Табл. 1 Эффективность автоматической сегментации в %

| Список | Число диад | Число отсутствующих сегментаций | Среднее значение | Дисперсия |
|---|---|---|---|---|
| 1 | 112 | 382 | 3,41 | 20,33 |
| 2 | 109 | 516 | 4.61 | 47.19 |
| 3 | 105 | 671 | 6.39 | 47.23 |
| 4 | 118 | 753 | 6.38 | 52.64 |

Табл. 2 Отсутствующие сегментации в списках логатомов

| Тип диады | CV | VC | CC | VV | V̆C | CV̆ | V̆V | VV̆ | V̆V̆ |
|---|---|---|---|---|---|---|---|---|---|
| Число диад | 118 | 148 | 118 | 7 | 13 | 18 | 11 | 11 | 0 |
| % отсутств. сегментаций | 6.2 | 9.9 | 17.1 | 44.0 | 23.3 | 16.3 | 29.8 | 34.1 | - |

Табл. 3 Сегментация типов диад

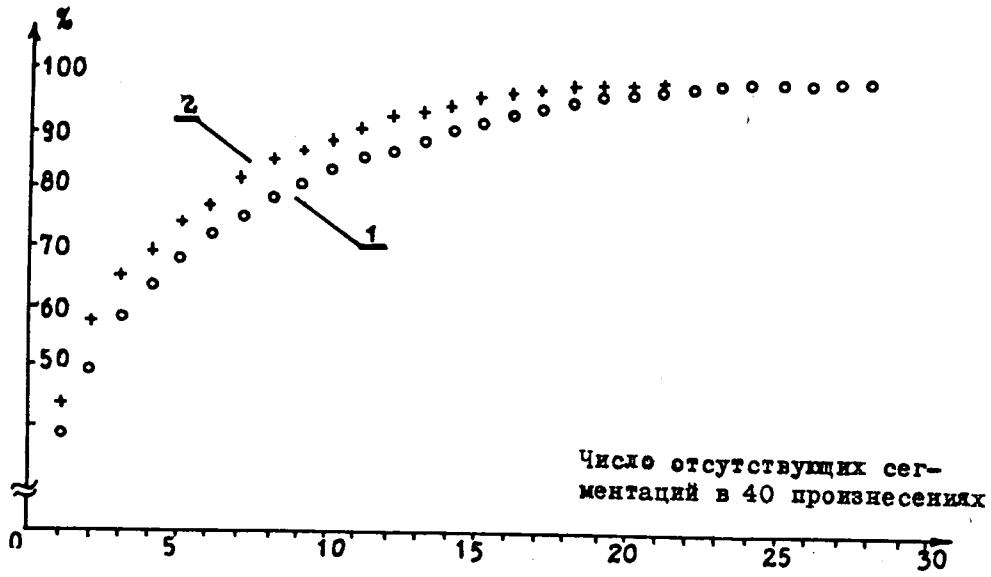


Рис. 1 Функция распределения отсутствующих сегментаций, 1 - без принятия во внимание и 2 - с принятием во внимание частотности диад

ЛИТЕРАТУРА

/1/ Домагала, П., Автоматизация процесса сегментации сигнала речи в аналого-цифровой системе, Работы Ин-та Основных Проблем Техники, №5/1984, Варшава.

/2/ Яссем, В., Лобач, П., Фонотактический анализ польского текста, Работы Ин-та Осн. Пробл. Техн., №63/1971, Варшава.

/3/ Лобач, П., Фонетико-лексикальные взаимодействия в восприятии речи, Изд. Ун-та А. Мицкевича, Познань, 1985.