О. Б. Зельманский, А. Г. Давыдов

ПОСТРОЕНИЕ ПРОГРАММНОГО МОДУЛЯ СЕГМЕНТАЦИИ РЕЧИ НА ОСНОВЕ АНАЛИЗА ИЗМЕНЕНИЯ СПЕКТРА

Введение

Ввиду того что речевой сигнал представляет собой последовательность квазистационарных участков, соответствующих голосовым и шумовым фонемам, между которыми располагаются участки с быстрым изменением параметров сигнала, соответствующие переходам между фонемами, сегментация речи представляет собой процесс разбиения речи на участки однородных колебаний, соответствующие разным типам фонем: гласноподобные, назальные, фрикативные, смычные. Задача сегментации речи является основной при построении систем распознавания и синтеза речи, а также систем активной защиты информации, осуществляющих синтез речеподобных помех на основе базы фонетических единиц, выделяемых непосредственно из маскируемого речевого сигнала в режиме реального времени [1].

Таким образом, существует необходимость в разработке модуля сегментации речевого сигнала с целью его последующего использования в системах активной акустической маскировки, формирование помехи в которых осуществляется из речи участников защищаемых конфиденциальных переговоров.

Описание метода

Следует выделить два подхода к решению задачи сегментации речи: разбиение речевого сигнала на фиксированные участки с последующим определением их принадлежности к тем или иным фонемам и обнаружение границ между фонемами с последующим распознаванием выделенной фонемы. В большинстве современных систем распознавания речи преобладает первый подход. Известно несколько видов алгоритмов сегментации. Первый характеризуется тем, что заранее известна последовательность фонем анализируемого речевого участка [2], второй основывается на изменениях акустических характеристик сигнала, априорная информация при этом не используется [3]. Третий подход заключается как в исследовании акустических параметров сигнала, так и в использовании априорной информации [4]. Поскольку предполагается использовать автоматическую сегментацию, то целесообразно применить второй подход, использующий только общие характеристики речевого сигнала, такие, например, как функция изменения спектра.

В соответствии с [5], функция изменения спектра сигнала определяется как мера корреляции между последовательными окнами анализируемого сигнала и может рассчитываться исходя из расстояния между параметрическими векторами x и y соответствующих окон сигнала несколькими способами.

1. Евклидово расстояние представляет собой геометрическое расстояние в многомерном пространстве и вычисляется по формуле:

$$E = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \ . \qquad (1)$$

В тех случаях, когда наиболее отличающимся векторам необходимо придать большее значение расстояния, возможно использование квадрата евклидова расстояния. Кроме того, когда каждому параметру в параметрическом векторе назначается свой удельный вес w_i, пропорциональный степени важности параметра, используется взвешенное евклидово расстояние:

$$E = \sqrt{\sum_{i=1}^{N} w_i (x_i - y_i)^2} \ . \qquad (2)$$

2. Нормированное евклидово расстояние для любой пары векторов всегда больше 0 и меньше или равно 1:

$$E = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i}{n_i} - \frac{y_i}{n_i}\right)^2}\ , \tag{3}$$

где n_i — норма i-го параметра, которая представляет собой разность между максимальным и минимальным значениями параметра анализируемых векторов.

3. Расстояние «городских кварталов» («сити-блок» расстояние, расстояние Хемминга в случае бинарных данных) представляет собой сумму модулей разности между векторами по каждому параметру:

$$D = \sqrt{\sum_{i=1}^{N} | x_i - y_i |}\ . \tag{4}$$

Для данного вида расстояния влияние отдельных больших разностей по сравнению с евклидовым расстоянием уменьшается, применяется, когда необходимо определить, насколько различаются вектора по каждому параметру.

4. Расстояние Чебышева принимает значение наибольшего модуля разности между значениями соответствующих параметров векторов:

$$D = \max | x_i - y_i |\ . \tag{5}$$

Это расстояние применяется, когда необходимо определить два вектора как сильно отличающиеся, если они различаются по какому-либо одному параметру.

5. Расстояние Минковского, или степенное расстояние, позволяет прогрессивно увеличить или уменьшить вес, относящийся к параметру, для которого соответствующие вектора сильно отличаются, имеет наиболее общий вид:

$$D = \left(\sum_{i=1}^{N} | x_i - y_i |^{p}\right)^{\frac{1}{r}}, \tag{6}$$

где ρ — параметр, ответственный за постепенное взвешивание разностей по отдельным параметрам, r — параметр, ответственный за прогрессивное взвешивание больших расстояний между векторами. При $\rho = r = 1$ данное расстояние соответствует расстоянию «городских кварталов», в случае $\rho = r = 2$ оно соответствует евклидову расстоянию.

6. Расстояние Махаланобиса принципиально отличается от предыдущих, поскольку учитывает корреляцию между параметрами и инвариантно масштабу. Высчитывается по формуле:

$$D = \sqrt{(x_i - y_i)^T S^{-1} (x_i - y_i)}\ , \tag{7}$$

где S^{-1} — обращенная ковариационная матрица.

Расстояние Махаланобиса может интерпретироваться как расстояние между заданной точкой и центром масс, деленное на ширину эллипсоида в направлении заданной точки.

7. Косинусное расстояние определяется как единица минус косинус от угла между параметрическими векторами:

$$D = 1 - \frac{xy^T}{(x^T x)^{\frac{1}{2}}} (y^T y)^{\frac{1}{2}}\ . \tag{8}$$

8. Корреляционное расстояние определяется как единица минус выборочный коэффициент корреляции между значениями параметров векторов.

$$D = 1 - \frac{(x-\bar{x})(y-\bar{y})^T}{[(x-\bar{x})(x-\bar{x})^T]^{\frac{1}{2}}[(y-\bar{y})(y-\bar{y})^T]^{\frac{1}{2}}}, \quad (9)$$

где $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

Результаты применения вышеописанных расстояний в качестве функции изменения спектра при сегментировании речевых сигналов приведены в таблице 1.

Таблица 1. Результаты использования различных метрик в качестве функции изменения спектра

	Евклидово расстояние	Нормированное евклидово расстояние	Расстояние Махаланобиса	Расстояние «городских кварталов»	Косинусное расстояние	Корреляционное расстояние	Расстояние Минковского с метрикой 3	Расстояние Минковского с метрикой 4
Количество лишних границ, %	18,2	16,8	21,8	18,8	20,9	22,9	18,1	17,5
Количество пропущенных границ, %	30,9	31,0	33,9	32,5	35,2	35,9	31,3	27,9
Точность обнаружения границ, %	98,6	98,9	98,7	99,0	98,9	98,7	98,8	98,7

В [6] приводится несколько вариантов определения функции изменения спектра, успешная реализация которых описана в [5]. Однако этим реализациям присущ один недостаток — они чувствительны к шуму, а именно из-за шума на всех участках сигнала возникают ложные пики функции изменения спектра.

Одним из подходов к сегментации речевых сигналов при реализации средств защиты речевой информации на основе генерирования речеподобного сигнала [7] является использование в качестве параметров сигнала его кепстральных коэффициентов. Вектор кепстральных коэффициентов может быть получен с помощью обратного *ДПФ* от логарифма амплитуды спектра, полученного с помощью прямого *ДПФ* (Рис. 1).

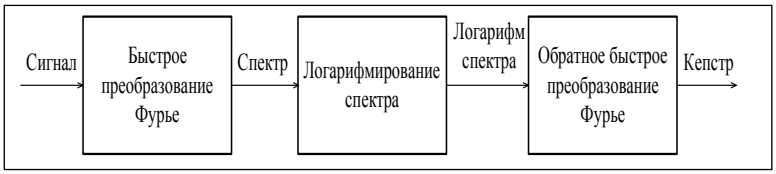

Рис. 1. Схема вычисления кепстра сигнала

К наиболее эффективным методам, обеспечивающим высокую точность сегментирования и основанным на кепстральном анализе сигнала, относятся [8]: метод кепстральных коэффициентов линейного предсказания (LPCC – Linear Prediction Cepstra Coefficients) [9], метод линейных спектральных частот (LSF – Linear Spectral Frequencies) [10, 11], метод коэффициентов перцептивного линейного предсказания (PLP – Perceptual Linear Prediction) и робастный PLP (RASTA-PLP – RelAtive SpecTrAl PLP) [12], метод кепстральных коэффициентов тональной частоты на шкале мел (MFCC – Mel-Frequency Cepstra Coefficients) [13].

На сегодняшний день наибольший интерес представляет алгоритм LSF, который в отличие от алгоритма LPCC основывается не на коэффициентах линейного предсказания модели голосового тракта [9], а на эквивалентных им линейных спектральных частотах (корнях), которые достаточно просто рассчитываются, экономно представляются и являются помехоустойчивыми [10]. Корни в общем случае могут быть получены в результате решения двух уравнений [11]:

$$\begin{cases} \mathrm{Re}\left\{z^{R} A_p(z)\right\}_{z=e^{i\tilde{w}}} = 0, \\ \mathrm{Im}\left\{z^{R} A_p(z)\right\}_{z=e^{i\tilde{w}}} = 0, \end{cases} \text{при } R \geq \frac{p}{2}\ , \tag{10}$$

где $A_p(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}, \tilde{w} = wT$.

При обозначении:

$$x = z + z^{-1} = 2\cos\tilde{w}, \tag{11}$$

линейные спектральные корни вычисляются по формуле:

$$\tilde{w}_n = \arccos x_n . \tag{12}$$

В зависимости от R корни могут рассчитываться по-разному. Так, при $R = \frac{p}{2}$ получается минимально возможный порядок уравнений, при $R = \frac{p+1}{2}$ — классический случай, подробно изложенный в [10], при $R > \frac{p+1}{2}$ — избыточное число параметров, а при $R = p$ достаточно решить только одно уравнение порядка N, чтобы по его корням найти все коэффициенты исходного многочлена [11].

Наибольшая информация о сигнале содержится в первых шести кепстральных коэффициентах, включение остальных коэффициентов зависит от типа решаемой задачи и особенностей произношения диктора.

Таким образом, в результате проведенных исследований для построения модуля сегментации речи был использован основанный на кепстральном анализе сигнала алгоритм LSF. Для расчета функции изменения спектра было использовано расстояние Минковского с метрикой, равной четырем, поскольку применение этого расстояния обеспечивает наиболее оптимальное соотношение между количеством лишних границ, пропущенных границ и точностью их обнаружения (таблица 1).

Программная реализация метода

Для реализации модуля сегментации речи была выбрана среда символьной математики MATLAB. Все математические операции, выполняемые в процессе работы программы, формализованы на языке системы MATLAB непосредственно в программном коде. Данная реализация обеспечивает переносимость полученного программного модуля.

В ходе системного проектирования была разработана структурная схема модуля сегментации речи, в основе которой лежит описанный выше алгоритм анализа изменения спектра сигнала. Данная схема представлена на рис. 2.

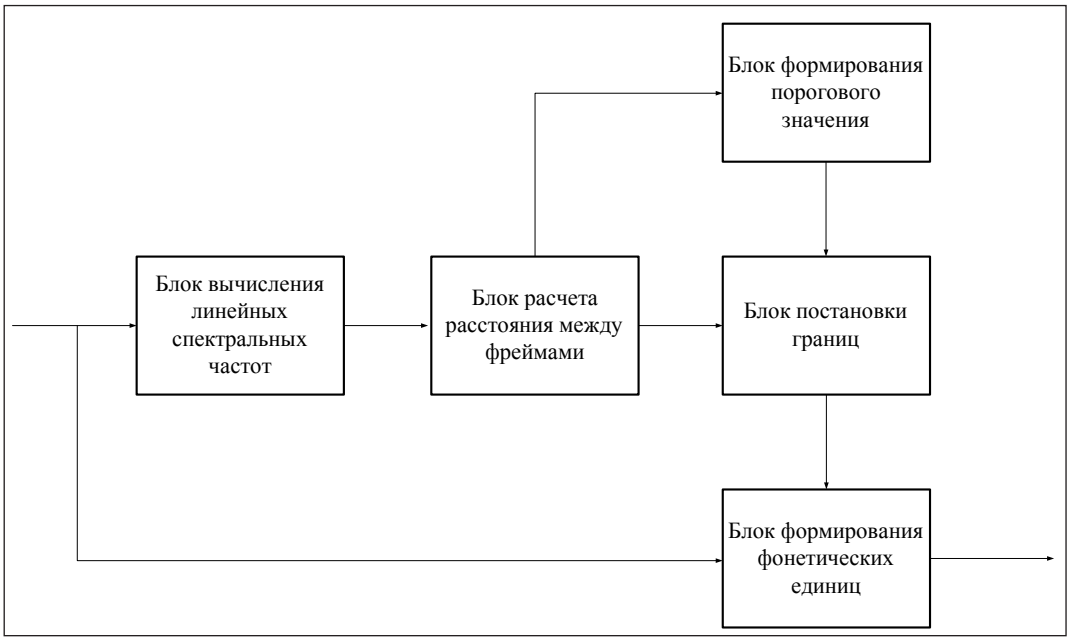

Рис. 2. Структурная схема модуля сегментации речи

На вход данного модуля поступает массив речевых участков, представляющий собой последовательность временных фреймов обрабатываемого речевого сигнала, выделенных модулем детектирования речи. Длительность каждого временного фрейма составляет 50 мс. На выходе модуля сегментации речевого сигнала формируется массив фонетических единиц речи. В основу модуля положена идея анализа изменения спектра сигнала. Для его реализации требовалось решить следующие задачи.

1. Для каждого временного фрейма требуется рассчитать кепстральную характеристику, для чего необходимо вычислить эквивалентные коэффициентам линейного предсказания линейные спектральные частоты (LSF). Для этого был разработан блок вычисления линейных спектральных частот. Данный блок анализирует сигнал в текущем временном фрейме в частотной области и выполняет следующие функции:

- расчет параметров модели голосового тракта;
- расчет линейных спектральных частот.

На выход блока поступает вектор линейных спектральных частот.

2. Далее выполняется расчет расстояний между соседними временными фреймами исходя из соответствующих им векторов рассчитанных линейных спектральных частот. Для этой цели был разработан блок расчета расстояния между фреймами. Функцией данного блока является расчет расстояния Минковского для последовательных фреймов. В качестве настроечного параметра данного блока выступает параметр расстояния Минковского, по умолчанию принимающий значение четыре. Выходом данного блока является значение расстояния между фреймами.

3. В соответствии с рассчитанными расстояниями между последовательными временными фреймами сигнала вычисляется пороговое значение расстояния, превышение которого определяет наличие границы между фонетическими единицами. В связи с этим был реализован блок формирования порогового значения. На вход данный блок получает значения расстояний между фреймами, на выход поступает пороговое значение расстояния. Настроечным параметром данного блока является отношение величины порогового значения расстояния к среднему значению расстояния между фреймами.

4. На основании расстояния между соседними временными фреймами и пороговым значением расстояния необходимо принять решение о постановке границы фонетической единицы. Для этой цели был создан блок постановки границ. В его функции входит:
- сравнение расстояния между фреймами с пороговым значением, которое адаптируется к анализируемому сигналу;
- принятие решения о наличии границы.

На вход блока поступают значения расстояний между фреймами, а также пороговое значение расстояния. На выходе формируется единица, в том случае если принято решение о постановке границы, ноль — в противном случае.

5. Заключительным этапом является формирование массива сегментов речевого сигнала, которые представляют собой фонетические единицы речи. Эту задачу выполняет блок формирования фонетических единиц. На его вход поступают временные фреймы сигнала и бинарные решения о наличии границы от блока постановки границ. Основной функцией данного блока является формирование массива фонетических единиц речи в соответствии с расставленными фонетическими границами. Полученный массив поступает на выход данного блока.

Заключение

Тестирование разработанного модуля сегментации речи производилось с помощью известного речевого корпуса TIMIT, исходно предназначенного для разработки и оценки систем автоматического распознавания речи. Акустико-фонетический корпус TIMIT американского варианта английского языка состоит из 2342 записей отдельных предложений 630 дикторов из 8 региональных диалектных зон США. Соотношение дикторов составляет около 70 % дикторов-мужчин и 30 % женщин.

Результаты тестирования приведены в таблице 2.

Таблица 2. Результаты тестирования модуля сегментации речи

Количество лишних границ, %	Количество пропущенных границ, %	Точность обнаружения границ, %
7,5	27,9	98,7

Как видно из таблицы 2, предлагаемый модуль сегментации речи позволяет обнаружить до 72 % границ между фонемами с точностью до 98 %. Число лишних границ при этом составляет 17,5 %. Для сравнения, известный алгоритм сегментирования RASTA-SVF обнаруживает 70,2 % границ с точностью 95,5 % [14].

Таким образом, разработанный модуль сегментации речи обеспечивает результаты, достаточные для того, чтобы применить данный модуль в системах синтеза речеподобных сигналов с целью формирования базы аллофонов дикторов, участвующих в защищаемых переговорах.

СПИСОК ЛИТЕРАТУРЫ:

1. *Зельманский О. Б., Давыдов А. Г.* Подходы к решению задачи сегментирования речи в рамках разработки генератора речеподобных сигналов // Современные проблемы радиотехники и телекоммуникаций: материалы МНТК. Севастополь, 2010. С. 386.
2. *Ganapathiraju A., Hamaker J., Picone J., Doddington G. R., Ordowski M.* Syllable-Based large vocabulary continuous speech recognition // IEEE Transactions on Speech and AudioProcessing. 2001. Vol. 9. № 4. P. 358–366.
3. *Kamakshi P., Nagarajan, Hema M.* Automatic segmentation of continuous speech using minimum phase group delay functions // Speech Communication. 2004. Vol. 42. P. 429–446.
4. *Цыплихин А. И., Сорокин В. Н.* Сегментация речи на кардинальные элементы // Информационные процессы. 2006. Том 6. № 3. С. 177–207.

5. *Flammia G. [et al.]* Segment based variable frame rate speech analysis and recognition using a spectral variation function // Interspeech 1992 — ICSLP: Proceedings of the second international conference on spoken language processing, Banff, Alberta, October 13–16, 1992. Banff, Alberta, Canada, 1992. P. 983–986.
6. A study on spectral variation functions applied to speech signals: final report / Aalborg University; Nouza, J. CPK. 1994. № 4678.
7. *Зельманский О. Б., Давыдов А. Г.* Система генерирования речеподобных сигналов для маскирования акустической информации // СВЧ-техника и телекоммуникационные технологии: материалы МНТК. Киев, 2010. С. 506–507.
8. *Зельманский О. Б., Давыдов А. Г.* Параметризация речевого сигнала в системах сегментации речи // Информационные системы и технологии (IST'2010): материалы VI Международной конф. Минск, 24–25 ноября 2010 г. / Науч.-технолог. ассоциация «Инфопарк»; редкол.: А. Н. Курбацкий [и др.]. Минск: А. Н. Вараксин, 2010. С. 163–166.
9. *Маркел Д. Д., Грэй А. Х.* Линейное предсказание речи. М.: Связь, 1980. — 308 с.
10. *Itakura F.* Line spectrum representation of linear predictor coefficients of speech signals // Acoustical society of America. 1975. Vol. 57. № 1. P. 77–86.
11. *Ланнэ А. А.* Новая теория линейных спектральных корней // Третья Международная конференция «Цифровая обработка сигналов и ее применение»: сборник трудов. Москва, 29 ноября — 1 декабря 2000 г. / РНТОРЭС им. А. С. Попова. М., 2000. С. 118–125.
12. *Калюжный А. Я., Семенов В. Ю.* Автоматическое определение пола диктора на основе гауссовых смесей // Акустический симпозиум «Консонанс-2009»: сборник тезисов конференции. Киев, 29 сентября — 1 октября 2009 г. / НАН Украины, Институт гидромеханики; редкол.: В. Т. Гринченко [и др.]. Киев, 2009. — 31 с.
13. *Граничин О. Н., Шалымов Д. С.* Решение задачи автоматического распознавания отдельных слов речи при помощи рандомизированного алгоритма стохастической аппроксимации // Нейрокомпьютеры: разработка, применение. 2009. № 3. С. 58–64.
14. *Petek B., Andersen O., Dalsgaard P.* On the robust automatic segmentation of spontaneous speech // Proc. ICSLP. 1996. P. 913–916.