**IX Всероссийская конференция молодых ученых
по математическому моделированию
и информационным технологиям**

28–30 октября 2008 года, г. Кемерово

Уменьшение информационной избыточности речевого сигнала методом амплитудной фильтрации

Н.В. Бобров

Московский государственный лингвистический университет
Центр фундаментального и прикладного речеведения
119034, Москва, ул. Остоженка, 38, корп. Б, лаб. 204

Введение

Большинство современных методов анализа и обработки речевого сигнала основаны на исследовании его частотного спектра. Их плюсом является принципиальная аналогия с работой человеческого слуха, минусом – неизбежно высокая ресурсоемкость. Иногда последняя выступает как фактор, ограничивающий возможности разработки устройств передачи и анализа речи.

Альтернативу составляют методы, работающие во временной области. Входными данными для них являются непосредственно ординаты последовательных точек осциллограммы звукового давления. В 1960–1970 гг. они разрабатывались наравне с методами, использующими информацию ее частотного спектра, однако впоследствии, с увеличением доступности мощных ЭВМ, это направление отошло на второй план.

В связи с этим стоит упомянуть еще одну тенденцию, сыгравшую немаловажную роль в развитии речеведения во второй половине XX века, – применение математического моделирования при описании звучащей речи. Появление акустической теории речеобразования Г. Фанта [Фант 1964] доказало состоятельность этого подхода и в то же время обнаружило колоссальные препятствия на пути его реализации. Модели, адекватно описывающие процессы речепроизводства и речевосприятия, сложны, включают большое количество трудно поддающихся измерению параметров (без которых они не могут претендовать на универсальность) и неудобны для программной реализации на ЭВМ. Последнее обстоятельство объясняется тем, что объектами моделирования в этих случаях выступают непрерывные процессы, в то время как современные вычислительные машины – дискретны.

Это противоречие остается неустранимым до сих пор, но его можно обойти, пойдя по пути одного из двух возможных

компромиссов. Первый заключается в использовании вычислительных систем с очень высокой производительностью, позволяющих «стереть» грань между дискретными и непрерывными процессами (на самом деле – лишь вывести ее за пределы субъективного человеческого восприятия). Цена этого решения измеряется издержками, связанными с обеспечением необходимых вычислительных мощностей. Второй компромисс состоит в том, чтобы с самого начала брать в качестве объекта для изучения и моделирования именно дискретное, а не непрерывное представление речевого сигнала. В этом случае ценой является априорный отказ от переноса полученных результатов на непрерывные физические процессы речевосприятия и речепроизводства при сохранении полного диапазона возможностей исследования дискретных речевых фонограмм и минимальных затратах вычислительных ресурсов. Выбор того или иного компромисса в каждой отдельной ситуации определяется постановкой задачи.

Предлагаемый нами метод амплитудной фильтрации речевого сигнала принадлежит к множеству решений, идущих по второму пути. Он основан на удалении из речевой волны пиков малой магнитуды.

Магнитуда пиков и сущность амплитудной фильтрации

Речевая волна в дискретном представлении является ломаной линией сложной формы (см. рис. 1). Каждый участок этой линии содержит конечное число узлов (вершин), которое можно оценить сверху по формуле:

$$N \le F_{\partial} t \,, \qquad (1)$$

где F_{∂} – частота дискретизации при аналого-цифровом преобразовании, а t – длительность отрезка речевого сигнала, соответствующего данному участку ломаной.

Заметим, что указанное максимальное значение достигается лишь тогда, когда в составе речевой волны присутствуют колебания, частота которых равна половине частоты дискретизации (частоте Найквиста для данного образца сигнала), и вообще число узлов ломаной находится в прямой зависимости от верхней границы спектра речевого сигнала на рассматриваемом участке.

Хотя не все узлы ломаной линии соответствуют экстремумам исходной кривой звукового давления, технически все они могут рассматриваться как пики. Каждый пик характеризуется магнитудой, которая в данном случае определяется как его относительная «высота», то есть кратчайшее расстояние от вершины до линии, соединяющей два соседних пика.

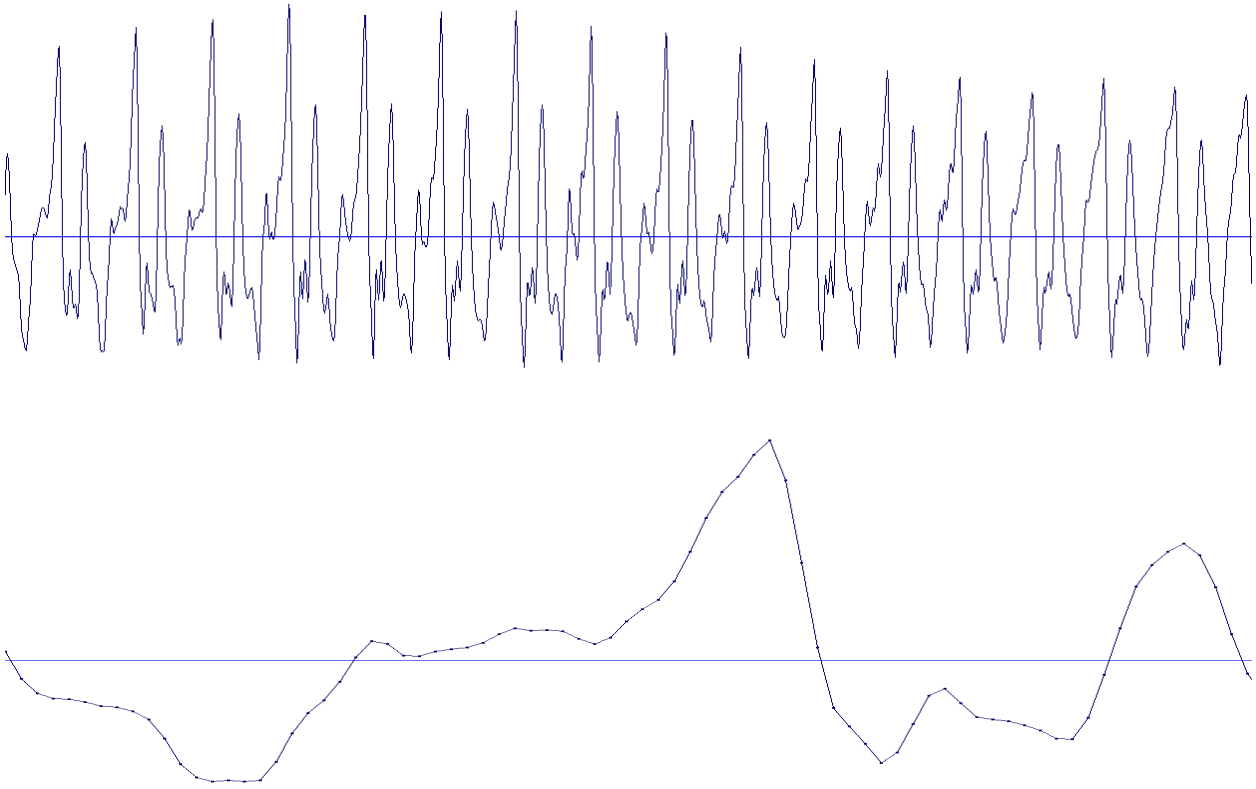

Рис. 1. Фрагменты речевой волны (гласный [а] в слове «десятая», женский голос). На нижнем примере видно, что после оцифровки (дискретизации) речевая волна представляет собой не кривую, а ломаную линию с большим количеством узлов.

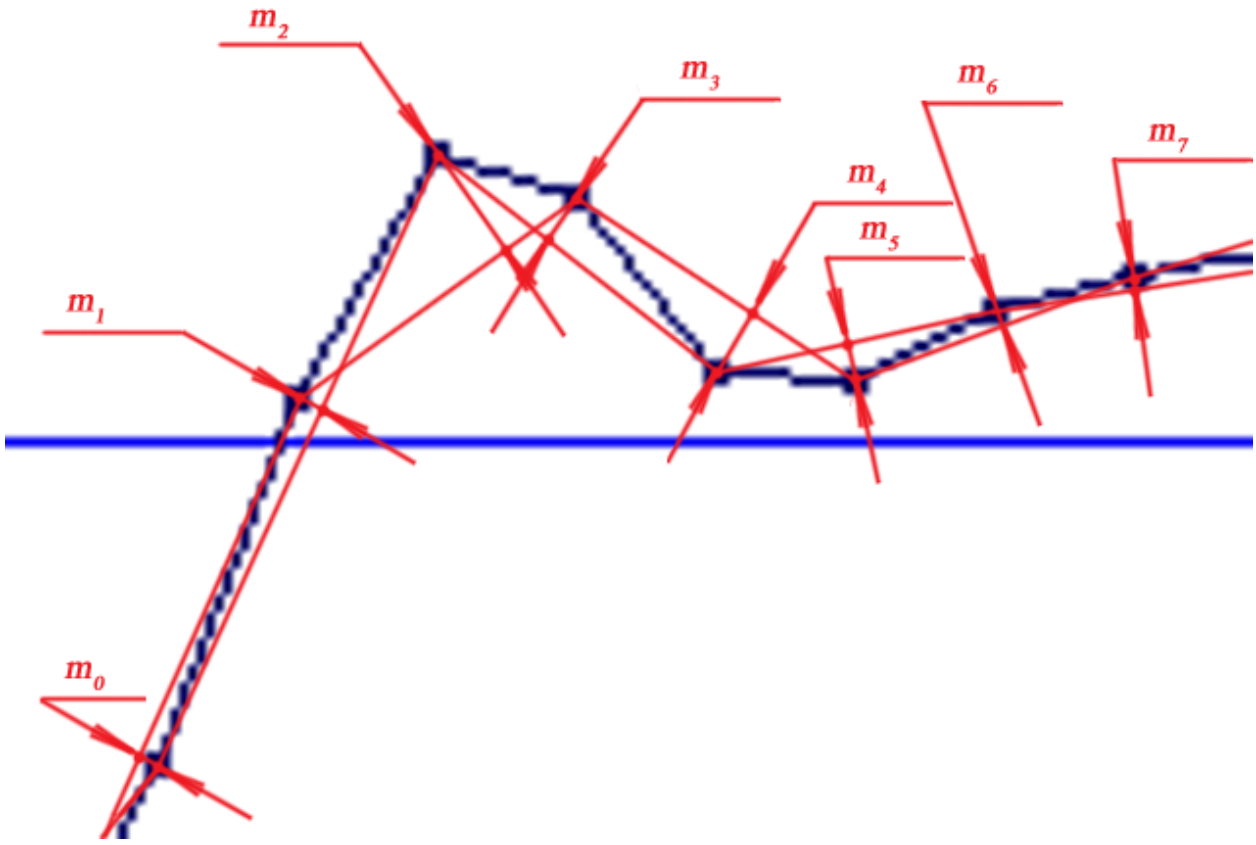

Рис. 2. Измерение магнитуд пиков речевой волны в дискретном представлении.

По рис. 2 видно, что магнитуда пиков речевого сигнала может сильно различаться. Кроме того, можно заметить, что существует некое пороговое значение магнитуды, при котором единичный пик начинает оказывать определяющее влияние на форму речевой волны. Это пороговое значение можно оценить экспериментально, последовательно исключая из речевой волны пики, амплитуда которых ниже заданного порога, и наблюдая за качеством (разборчивостью) результирующего сигнала. В данном примере пороговое значение магнитуды, вероятно, лежит между m_1 и m_5.

Удаление из речевой волны пиков малой магнитуды (т.е. тех, чья магнитуда оказалась ниже экспериментально определенного порогового значения) приводит образованию ломаной линии более простой формы, которая, тем не менее, несет в себе количество информации, достаточное как для визуального (по осциллограмме), так и для слухового опознавания звуков речи. Таким образом, удаленной оказывается главным образом именно избыточная информация, содержащаяся в речевом сигнале. Это и составляет идейную основу метода амплитудной фильтрации.

Несущественность пиков малой магнитуды для восприятия речи объясняется двумя особенностями, которые отличают речь от других акустических сигналов.

Первая из них заключается в характере распределения энергии в спектре речевого сигнала. На рис. 3 показан спектральный срез, полученный на трех периодах стационарного участка ударного гласного [а] в слове «десятая». Видно, что интенсивность колебаний, выраженная в децибелах, обратно пропорциональна частоте. Это означает, что их амплитуда, выраженная в абсолютных единицах, находится в обратной экспоненциальной зависимости от частоты. Данное наблюдение справедливо для гласных и сонорных согласных звуков, а также (с некоторыми оговорками) для звонких фрикативных. Отсюда следует, что для указанных классов звуков пики малой магнитуды в речевом сигнале представляют, как правило, именно высокочастотную часть спектра.

Вторая особенность речи заключена в механизме ее восприятия и состоит в том, что наиболее важными для разборчивости звуков речи, относящихся к вышеперечисленным классам, являются колебания с частотами от 300 до 4000 Гц, то есть как раз те, что в наименьшей степени затрагиваются при удалении из речевой волны пиков малой магнитуды.

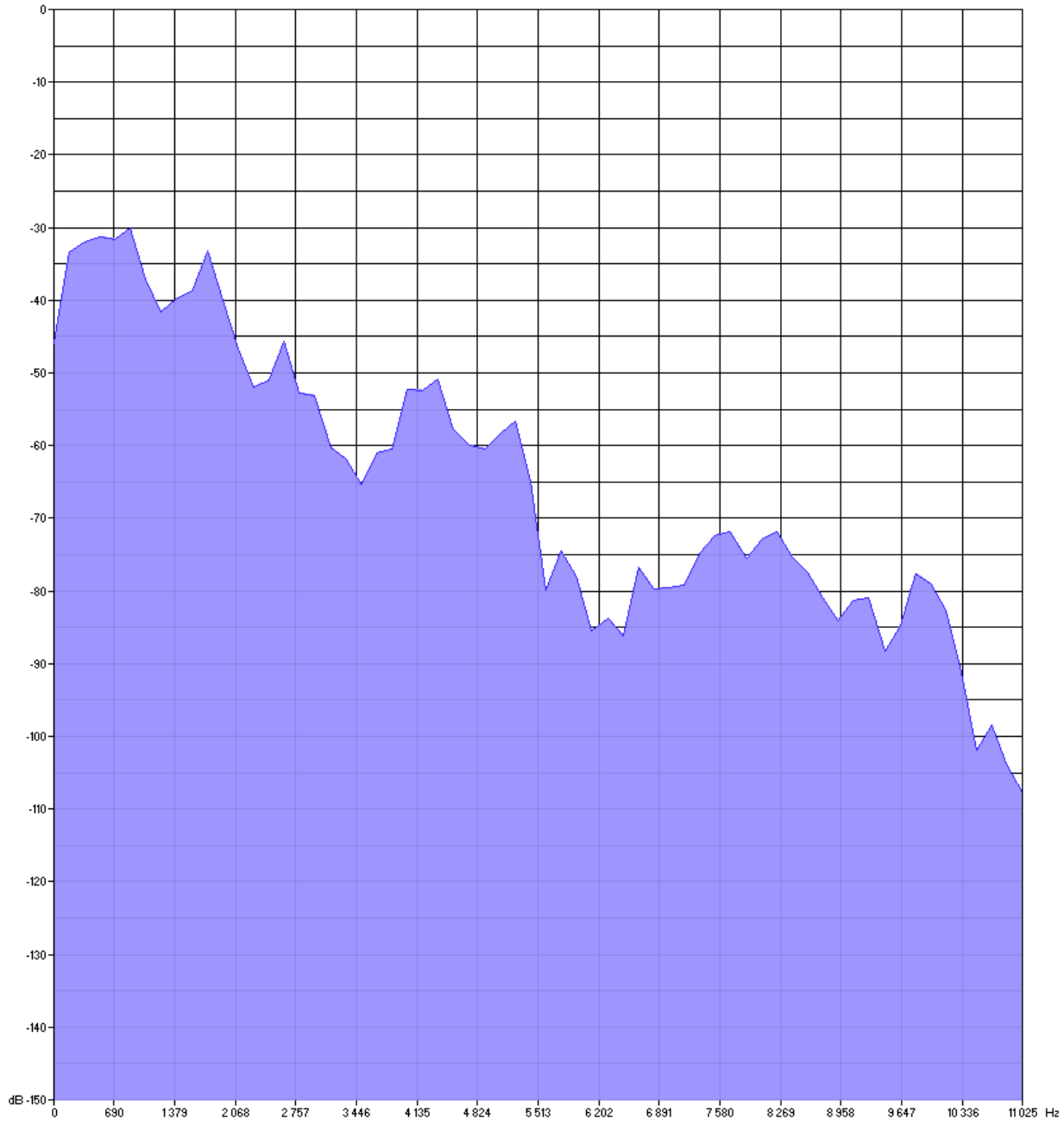

Рис. 3. Спектральный срез, полученный на трех периодах стационарного участка ударного гласного [а] в слове «десятая».

В случае смычных и фрикативных звуков удаление пиков малой магнитуды приводит к исчезновению из спектра некоторых компонентов (главным образом – находящихся на краях или вовсе за пределами их зон усиления), что также лишь незначительно влияет на разборчивость речи.

Можно сказать, что удаление пиков малой магнитуды в некотором смысле равнозначно понижению частоты дискретизации, избирательно применяемому к участкам речевого сигнала, соответствующим различным классам звуков. Коренное отличие предлагаемого подхода от избирательного понижения частоты дискретизации с использованием обычных предназначенных для этого

методов состоит в том, что в данном случае определение классов звуков и границ соответствующих им участков в речевом сигнале происходит само собой: эта информация является, по сути дела, побочным продуктом обработки сигнала.

Программная реализация метода амплитудной фильтрации

Несмотря на идейную простоту, вычислительная реализация описанного выше метода амплитудной (или, точнее, магнитудной) фильтрации является достаточно сложной задачей. Дело в том, что, если запрограммировать процесс именно так, как он был представлен в предыдущем разделе, получится, что сначала нужно определить, какие точки осциллограммы являются пиками, а затем для каждого пика рассчитать магнитуду – кратчайшее расстояние от него до прямой, соединяющей два соседних пика. Ресурсоемкость метода АФ при такой реализации окажется сопоставимой с ресурсоемкостью обычных спектральных методов анализа и обработки речевого сигнала или, может быть, будет даже превышать ее.

В целях минимизации ресурсозатрат нами был разработан альтернативный алгоритм, позволяющий получить тот же результат (с незначительной потерей точности) при минимальном объеме вычислений. Суть его заключается в преобразовании исходной сложной ломаной линии осциллограммы в сумму более простых ломаных линий и последующем исключении из этой суммы наименее значимых слагаемых.

Обработка потока данных ведется в пооконном режиме. Длина окна анализа (n) подбирается произвольно. Как правило, целесообразно использовать окно, в котором умещается около 10 мс речи (это соответствует обычной минимальной длительности речевого сегмента).

На каждом шаге цикла обработки данных ломаная в окне анализа представляется в виде суммы двух ломаных: упрощенной и остаточной.

Упрощенная ломаная получается из исходной путем целочисленного деления и умножения ординат ее точек на некоторую постоянную величину (k). Эта величина задается в начале выполнения программы как параметр, определяющий чувствительность преобразования, и может подвергаться корректировке (нормировке на среднюю амплитуду сигнала) перед началом обработки данных в окне анализа. Для ускорения работы программы вместо операций целочисленного деления и умножения возможно использование операций битового сдвига вправо и влево на величину k_2, равную целочисленному двоичному логарифму k.

Результатом данного преобразования является ступенчатая ломаная линия. Чтобы получить искомую упрощенную ломаную, необходимо заменить ее горизонтальные и вертикальные участки наклонными, соединив прямолинейными отрезками соседние вершины.

Остаточная ломаная (или остаточный сигнал следующего уровня) получается вычитанием только что построенной упрощенной ломаной из исходной. Она используется в качестве исходной на следующем шаге цикла.

Цикл продолжает работу до тех пор, пока среднее значение амплитуды остаточного сигнала не станет меньше некоторого уровня q (эта величина задается как параметр, определяющий точность преобразования).

Ход обработки данных в окне анализа показан на рис. 4.

По окончании работы цикла упрощенные ломаные всех уровней суммируются. Результатом является ломаная линия, представляющая собой дискретную осциллограмму речевого сигнала с уменьшенной информационной избыточностью (рис. 5).

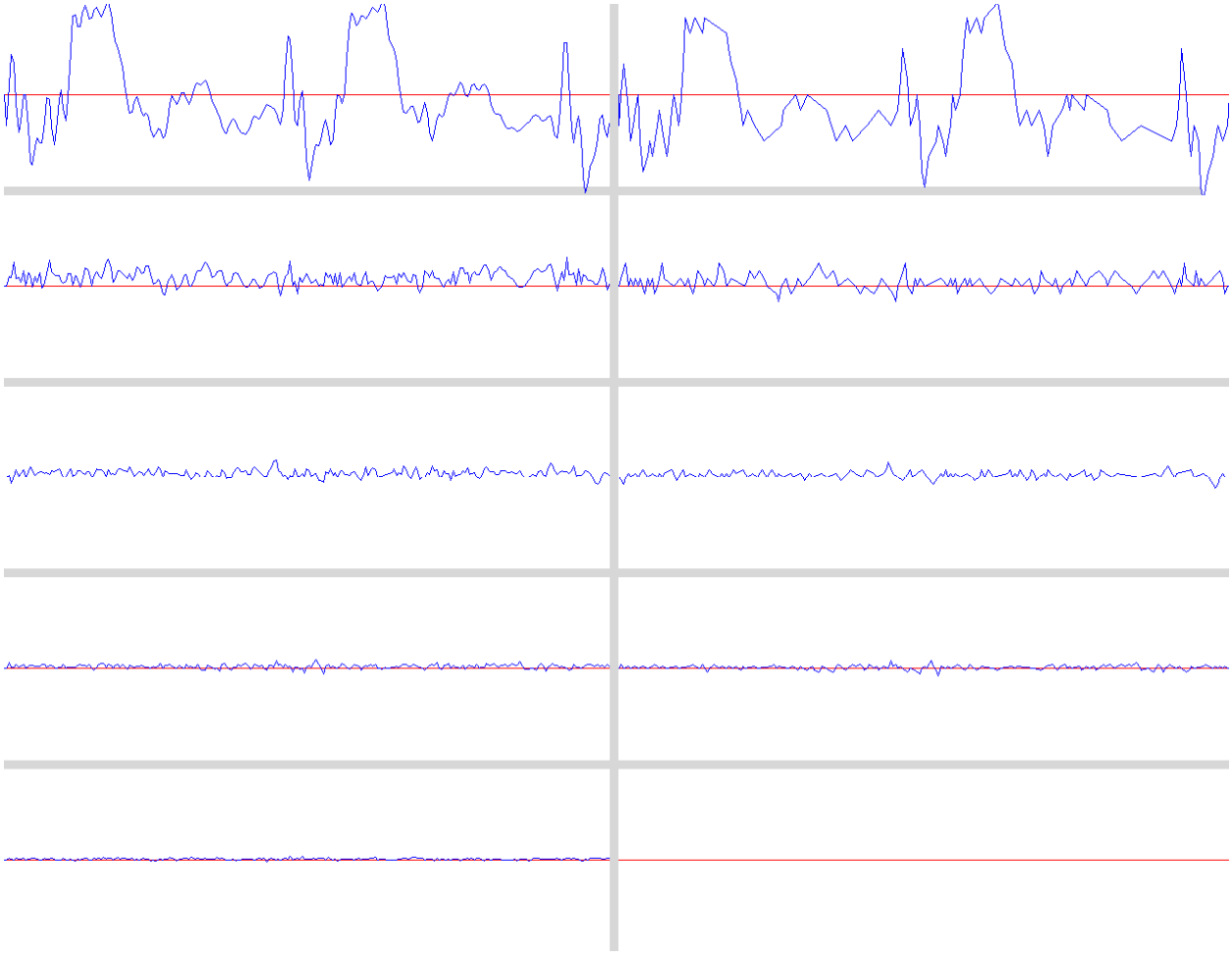

Рис. 4. Преобразование речевой волны в сумму ломаных линий. Звук [а]. Слева – исходный сигнал и остаточные сигналы первого, второго и др. уровней, справа – упрощенные ломаные, полученные из этих сигналов.

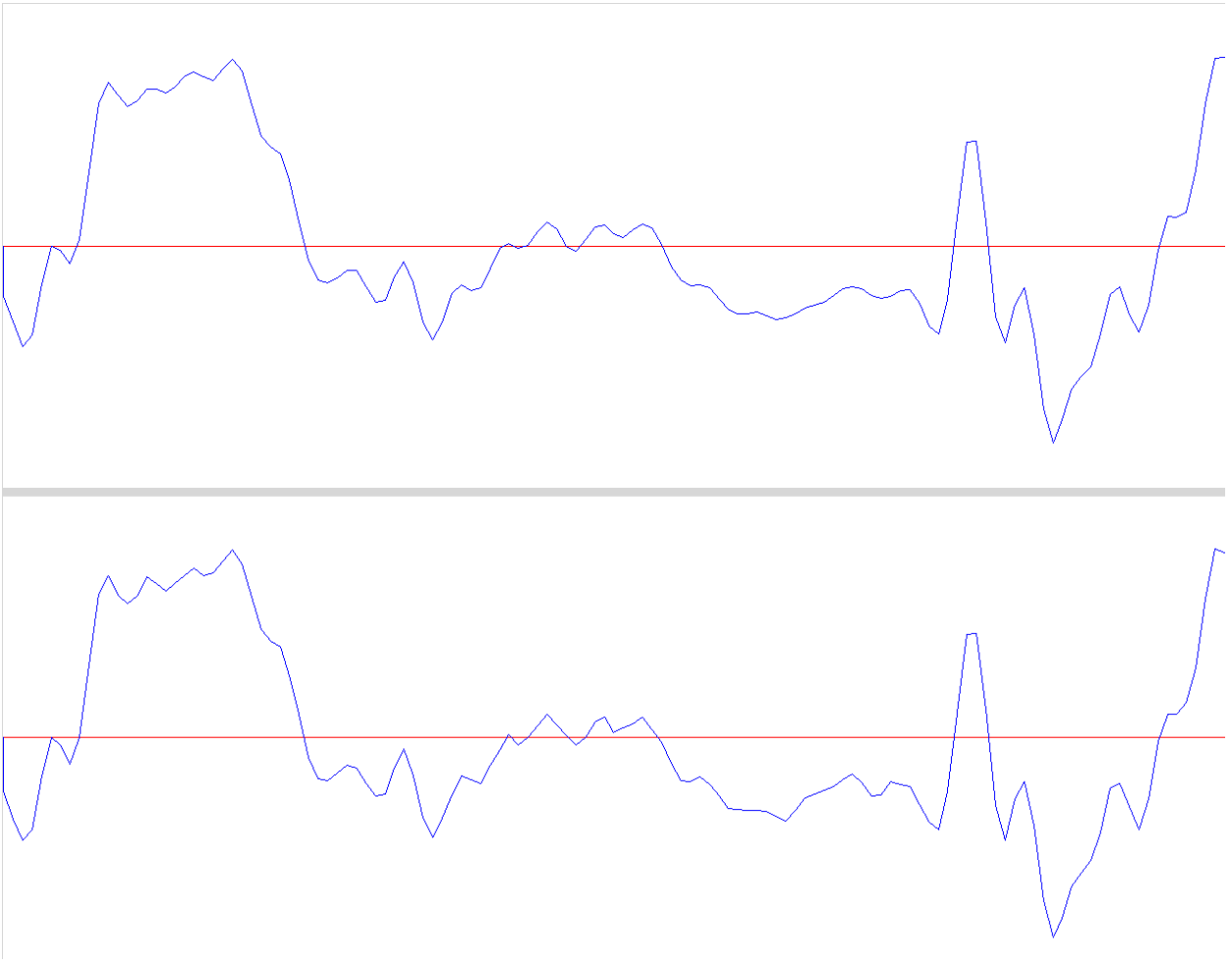

Рис. 5. Исходный (вверху) и результирующий сигналы. Гласный [а]. Различия наиболее заметны в средней части приведенного фрагмента.

Заключение

Как показали предварительные эксперименты, рассмотренный метод амплитудной фильтрации может уменьшать информационную избыточность речевого сигнала приблизительно в полтора раза при сохранении разборчивости речи.

Ввиду исключительно малой ресурсоемкости описанного альтернативного алгоритма обработки речевого сигнала он может быть реализован в сверхмалых устройствах анализа, кодирования и передачи речевой информации.

В настоящее время нами создан действующий прототип программной библиотеки для обработки речевого сигнала методом амплитудной фильтрации, ведутся работы в следующих направлениях:

– поиск и обоснование алгоритма автоматического подбора параметров амплитудной фильтрации (n, k, q);

– поиск оптимального алгоритма сжатия результирующего сигнала (или результирующей совокупности упрощенных ломаных) с целью создания нового компрессированного формата для хранения и передачи речевой информации;

– исследование закономерных изменений в структуре упрощенных ломаных различных уровней, связанных с фонетическим качеством звуков речи, в интересах разработки системы автоматического распознавания речи;

– совершенствование и оптимизация программных модулей с целью увеличения устойчивости, скорости и степени автоматизации их работы, а также расширения диапазона предоставляемых возможностей.

Литература

1. Зиндер Л.Р. Общая фонетика. М.: Высшая школа, 1979. 312 с.

2. Потапова Р.К. Речь: коммуникация, информация, кибернетика. М.: Радио и связь, 1997. 528 с.

3. Фант, Г. Акустическая теория речеобразования. М.: Наука, 1964. 284 с.