УДК 000.684
к.т.н., доц. **Соловьев В.И.** (ВНУ им. В. Даля)

СЕГМЕНТАЦИЯ ЗВУКОВОГО СИГНАЛА В ЗАДАЧАХ ВЫЯВЛЕНИЯ МОНТАЖА В АУДИОФАЙЛАХ

Приведены результаты исследований путей реализации задачи сегментации речи при выявлении и локализации монтажа аудиофайлов. Излагаемое направление исследований базируется на выявлении участков аудиофайла с характерными признаками частоты основного тона голоса человека.
Ключевые слова: аудиофайл, сегментация, монтаж.

Приведені результати досліджень реалізації завдання сегментації мови при виявленні та локалізації монтажу аудіофайлів. Напрям досліджень, що викладається, базується на виявленні ділянок аудіофайлу з характерними ознаками частоти основного тону голосу людини.
Ключові слова: аудіофайл, сегментація, монтаж.

Results over of researches of ways of realization of task of segmentation of speech are brought at exposure and localization of editing of audiofiles. The expounded direction of researches is based on the exposure of audiofile areas with the characteristic signs of frequency of basic tone of voice of man.
Keywords: audiofile, segmentation, editing.

Вступление и постановка задачи. При решении различных задач цифровой обработки аудиофайлов, связанных с анализом речевых сигналов, весьма существенной является задача автоматической сегментации звукового сигнала [1,2,3]. Точность сегментации на паузы и речевую составляющую звукового сигнала весьма существенна в задачах построения математических моделей выявления монтажа аудиофайла по статистическим характеристикам пауз речи. Одной из первичных задач, которые необходимо решить, является разделение сигнала на паузы и речевую составляющую аудиосигнала [1,2]. Предложено множество решений этой задачи, которые достаточно эффективны в рамках различных конкретных исследований [1,2,3]. В большинстве методов анализа пауз в задачах выявления монтажа, как правило, рассчитываются различные статистические и спектральные характеристики звукового сигнала в паузах речи. Естественно, точность выделения границ речевых сегментов и пауз оказывает существенное влияние на точность оценки статистических характеристик пауз.

В представляемых ниже результатах исследования этого вопроса рассматривается направление методов разделения пауз и речевых сегментов, базирующееся на известных статистических и спектральных характеристиках сегментов речи [1,2,3,4].

На рис.1 и рис. 2 представлены типичные фрагменты, иллюстрирующие переход от пауз речи к собственно речевым составляющим аудиосигнала (появление – рис. 1, и окончание – рис. 2, речевой составляющей).

Важным методологическим моментом исследуемой задачи является существенная “ассиметрия” физических характеристик звукового сигнала при переходе от паузы к фрагменту речи (рис. 1) и, наоборот – от фрагмента речи к паузе (рис. 2). Переход к паузе часто сопровождается остаточными импульсациями звукового сигнала, что, естественно, необходимо учитывать при выделении границ пауз и речи для более точных статистических оценок характеристик пауз.

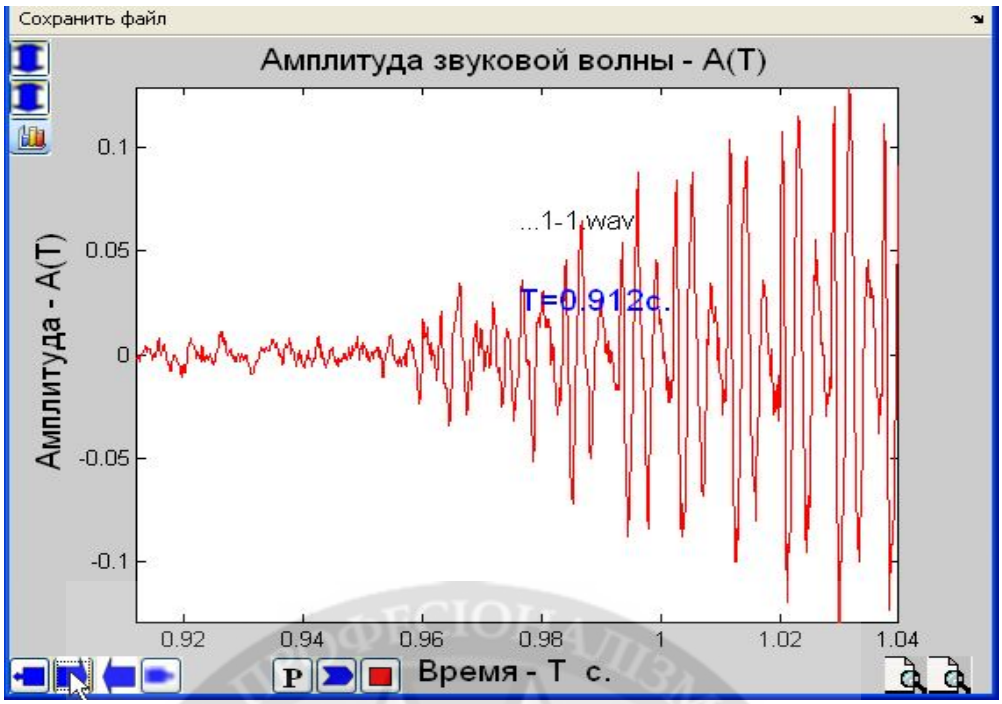

Рис. 1. Иллюстрация перехода от паузы к речевой составляющей аудиосигнала

Как показывают наши исследования, для выделения начала и окончания паузы необходима разработка различных методов разделения пауз и речевых составляющих звукового сигнала. Ниже рассматривается модель, которая эффективна при локализации момента начала сегмента речи (окончания паузы). Для эффективной локализации начала паузы, с точки зрения задач выявления монтажа аудиофайлов, требуются иные подходы, изложение которых будет рассматриваться в рамках других публикаций.

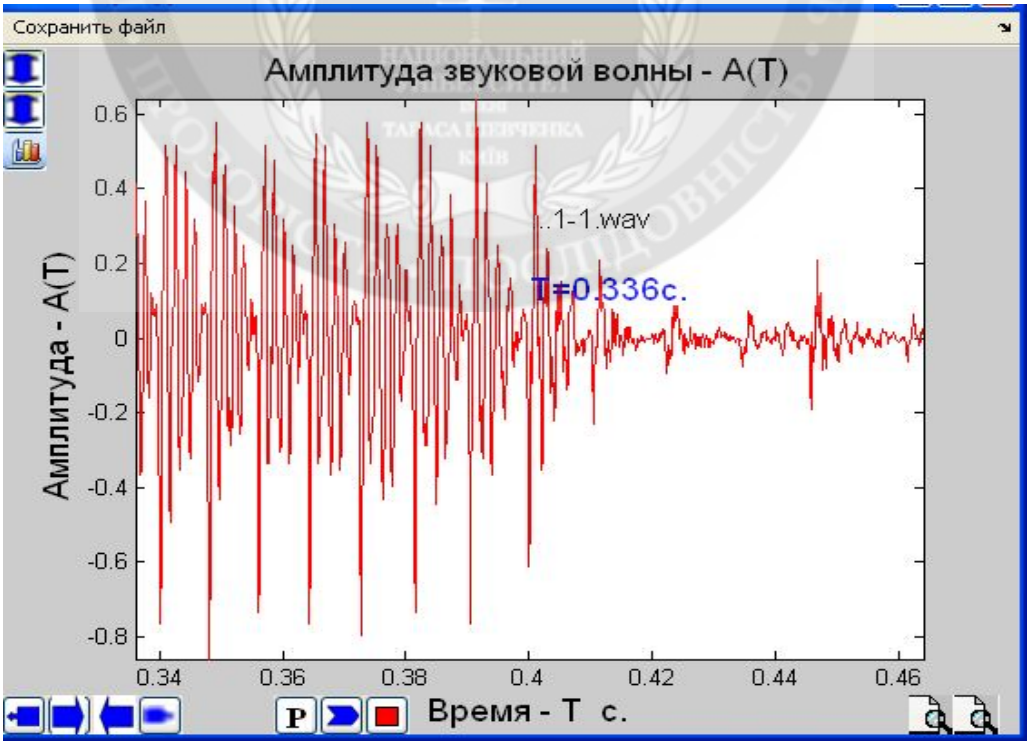

Рис. 2. Иллюстрация перехода от речевой составляющей аудиосигнала к паузе

Предлагаемая далее методология разделения пауз и речевых сегментов базируется на модели речевого сигнала в виде последовательности структур модулированных сигналов с несколькими частотными составляющими (возможно изменяющимися во времени). Характерной особенностью, по которой можно визуально разделить паузы и речевые фрагменты аудиосигнала, является появление структур модулированного сигнала с различными, явно выраженными, частотами (в области частот основного тона голоса человека). Рассмотрим возможность локализации границ окончания паузы и начала фрагмента речи по этому визуальному критерию.

В исследовании выборка аудиофайла сканировалась в перемещаемом временном окне с вычислением в окне локальных экстремумов аудиосигнала. Была поставлена задача выделения участков аудиофайла, на которых появляются локальные экстремумы аудиосигнала удовлетворяющие ряду требований характерных для речевых составляющих сигнала и отсутствующие в паузах.:

1. Интервал между локальными экстремумами, следующими последовательно друг за другом, равен – t с точностью Dt. Величина t находится в диапазоне времени, соответствующем величине, обратной диапазону частоты основного тона голоса. Величина Dt была выбрана в процессе многочисленных экспериментальных исследований в рамках рассматриваемой задачи.

2. Рассматривается не менее N=12 локальных экстремумов, следующих друг за другом. При этом эти экстремумы, удовлетворяющие первому условию, должны чередоваться по виду (минимум – максимум или наоборот).

3. Экстремумы либо увеличиваются по абсолютной величине, либо уменьшаются от первого к последнему (но не монотонно). Это условие моделирует начало нарастания или уменьшения (возможно с модуляциями) речевой составляющей аудиосигнала.

4. Среди соседних локальных максимумов в диапазоне интервалов времени, соответствующем частоте основного тона, выбирается абсолютный максимум.

5. Среди соседних локальных минимумов в диапазоне интервалов времени, соответствующем частоте основного тона, выбирается абсолютный минимум.

Излагаемая последовательность переработки информации по амплитудам звукового сигнала является схемой общего плана (без множества достаточно важных рабочих деталей алгоритма переработки информации).

На основе подхода, базирующегося на описании динамики локальных экстремумов, были проведены многочисленные исследования на различных аудиофайлах. Их цель – выявления оптимальных, с точки зрения точности локализации по времени, границ перехода между паузами и речевыми составляющими аудиосигнала. Варьировались параметры Dt и N.

Необходимо отметить следующее. Оптимальный выбор указанных выше параметров влияет не на точность локализации границы между паузами и фонемами, а на вероятности верного или ошибочного решения при автоматической локализации для обнаружения границы. Точность локализации границы при ее “безошибочном” обнаружении в данной методологии, как показывают экспериментальные исследования, порядка величины, обратной частоте дискретизации аудиосигнала. То есть, предельно возможная. Дело в том, что место локализации определяется в большинстве случаев по первому из последовательно следуемых экстремумов и равно временному отсчету соответствующему этому локальному экстремуму.

В связи с вышеуказанным, важен вопрос об ошибках первого и второго рода при принятии статистической гипотезы по вероятному обнаружению точки локализации конца паузы. Данный вопрос решался путем многочисленных экспериментов с варьированием параметров Dt и N для различных аудиофайлов (в формате просмотра wav).

Фактически эти вероятности зависят лишь от статистических характеристик аппаратуры аудиозаписи и фонового сигнала в процессе записи речевого аудиосигнала.

В частности, они полностью определяются вероятностью появления на паузах структур групп локальных экстремумов, удовлетворяющих рассматриваемой методике.

С точки зрения психофизиологии восприятия голоса, это эквивалентно тому, чтобы встретить в паузе случайно созданный голосовой фрагмент. Что является, казалось бы, совершенно невероятным событием. Однако, как показывают многочисленные психофизиологические исследования [5], человек начинает воспринимать “осмысленные” фонемы и их идентифицировать при достаточно большой их минимальной протяженности во времени. Это связано с психофизиологией восприятия звука. На рассматриваемых в исследовании временных отрезках (порядка 0,01 секунды), человек, как правило, не может идентифицировать по аудиосигналу на слух наличие начала фрагмента речи. В виду этого, на таких временных интервалах вполне вероятно встретить в паузах случайное сочетание локальных экстремумов, которое соответствует вышеуказанным требованиям.

Очевидно, что вероятность таких событий зависит не только от статистических характеристик пауз, но и от параметров методики.

На рис. 3 и рис. 4 представлены фрагменты исследования зависимости вероятности ложного обнаружения на паузах локализации начала фонем речи (рис. 4) и вероятности пропуска начала фонем речи (рис. 3), обобщенные по 50 аудиофайлам, записанным на различной аппаратуре аудиозаписи. При расчете вероятности пропуска начала фонем речи (рис. 3) в исследовании был принят критерий пропуска более двух локальных максимумов, которые при визуальном просмотре графика звукового сигнала характерны для начала сегмента речи. Этот визуальный критерий может показаться субъективным, однако, в процессе исследований при визуальном просмотре графиков аудиосигнала различными исследователями не было случаев двух различных мнений. Кроме того, данный субъективизм не оказывает никакого влияния на конкретную модель сегментации речевого сигнала.

Представленные зависимости получены при значении N=12, Dt – параметр, характеризующий разброс временных интервалов следования экстремумов (среднеквадратическое отклонение в алгоритме от величины среднего значения интервала следования экстремумов для 12 экстремумов, в % к среднему значению).

Приведенные иллюстрации фрагментов исследования характеризуют ошибки первого и второго рода в рамках рассматриваемой задачи.

Как видно из графиков зависимостей, ошибочное определение фрагмента начала речи имеет экстремум в районе Dt = 10%. Физически это обусловлено тем, что при уменьшении допуска по анализу следования интервалов экстремумов наступает момент, когда слишком “жесткие” условия на точность в алгоритме ухудшают вероятностные характеристики обнаружения.

Вероятность ложного обнаружения начала речи существенно зависит от частоты дискретизации звукового файла (при записи речи). Эта вероятность также существенно зависит, как показывают исследования, от качества аппаратуры аудиозаписи. В частности, при записи на ноутбуках часто в шумах записи присутствует помеха от сети на частоте 50 Гц и кратных ей частотах. Это существенно снижает эффективность работы в рамках рассматриваемой методики сегментации речи.

Для учета подобных “ложных” экстремумов необходим дополнительный анализ с дополнительными параметрами, которые отфильтровывают данный вариант шумов.

Вероятность ошибок как первого, так и второго рода в рассматриваемой задаче можно существенно снизить при увеличении параметра N. При этом, как показывают исследования, точность локализации границы при “безошибочном” обнаружении первых локальных экстремумов существенно возрастает.

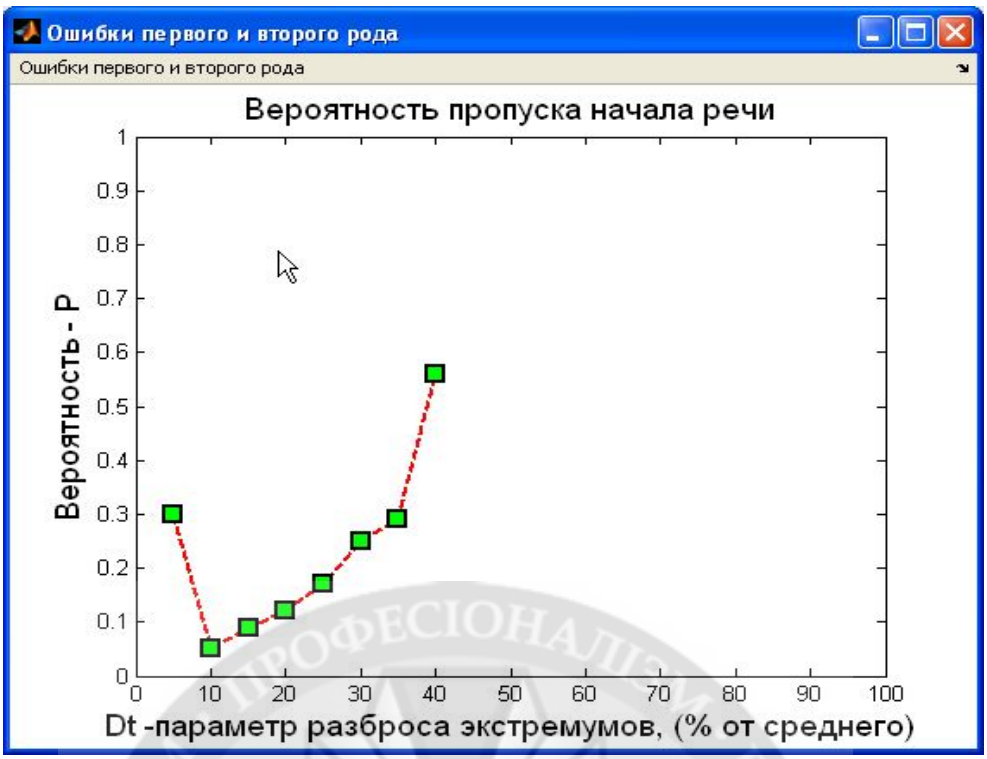

Рис. 3. Вероятность пропуска начала речи

Так, например, на рис. 5 и рис. 6 приведены графики вероятностных характеристик по ошибкам первого и второго рода, построенные по звуковым файлам, записанных на ноутбуке, с частотой дискретизации аудиосигнала – 44,1 кГц. Вероятность как пропуска начала речи, так и вероятность ложного обнаружения начала речи для этих файлов несколько выше при одном и том же значении параметра Dt.

Рис. 4. Вероятность ложного обнаружения начала речи

Рис. 5. Вероятность пропуска начала речи для аудиофайлов, записанных на ноутбуке

Рис. 6. Вероятность ложного обнаружения начала речи для аудиофайлов, записанных на ноутбуке

Выводы по результатам исследований:

1. Разработанное направление сегментации речевого сигнала и пауз речи обладает, при условии верного обнаружения фрагмента речи, максимально возможной точностью

локализации начала фрагмента речи – порядка времени обратному частоте дискретизации речевого сигнала.

2. Ошибки первого и второго рода в данной задаче (вероятность пропуска и начала речи и вероятность ложного обнаружения начала речи) зависят от ряда факторов, связанных как с характеристиками аппаратуры аудиозаписи, так и с параметрами алгоритма сегментации.

По величине эти ошибки являются вполне приемлемыми для решения ряда задач автоматической сегментации речи в задачах выявления монтажа аудиофайлов.

ЛИТЕРАТУРА:

1. Аграновский А.В. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов / А.В. Аграновский, Д.А. Леднов. – М: Радио и связь, 2004. – 164 с.

2. Солонина А.И. Основы цифровой обработки сигналов /А.И. Солонина, Д.А. Улахович, С.М. Арбузов, Е.Б. Соловьева. – С.-Пб.: БХВ-Петербург, 2-е изд., 2005. – 768 с.

3. Э. Айфичер. Цифровая обработка сигналов / Э. Айфичер, Б. Джервис. – М. – С.-Пб. – К.: Радио и связь, 2004. – 989 с.

4. С. Мала. Вейвлеты в обработке сигналов / С. Мала. – М.: Мир, 2005. – 671 с.

5. Александрова Ю. И. Психофизиология / Ю. И. Александрова. – М – С.-Пб.: Питер, 2006. – 463 с.

Рецензент: д.т.н., проф. Лєнков С.В.