Новый подход к определению границ речевого сигнала. Проблемы конца сигнала

Шелепов В.Ю., *доктор физико-математических наук, профессор*

Ниценко А.В., *программист*

В статье предлагается новый метод записи и выделения глухих взрывных согласных в конце слова, определения наличия звонкого согласного в конце постепенно затухающего сигнала. Инструменты: вариация с переменным верхним пределом по модулю 256, последовательные сглаживания.

• глухой взрывной • звонкий согласный в конце слова • вариация с переменным верхним пределом • сглаживание.

The content of the article is new method of recording and detachment toneless stop consonant in the end of the word, presence of voiced consonant in the end of gradually damped signal clarification. Instruments: variable upper boundary variation (module 256), smoothering.

• toneless stop consonant • voiced consonant in the end of the word • variable upper boundary variation • smoothering.

При записи речевого сигнала с целью распознавания речи требуется как можно более точное определение начала и конца записываемого речевого отрезка. Недопустимо, например, чтобы участок какого-либо низкоамплитудного звука речи оказался целиком «отрезанным». В связи с этим возникают так называемые проблемы конца сигнала. Первая проблема — запись и выделение в конце сигнала участков, отвечающих глухим взрывным звукам *[К], [П], [Т]*. Эти звуки являются паузообразными. Поскольку при их произнесении есть момент полного перекрытия голосового тракта, и при этом голосовые связки молчат, в сигнале появляется характерный паузообразный сегмент. В качестве примера приведём визуализацию сигнала, отвечающего слову «лапа» (рис. 1).

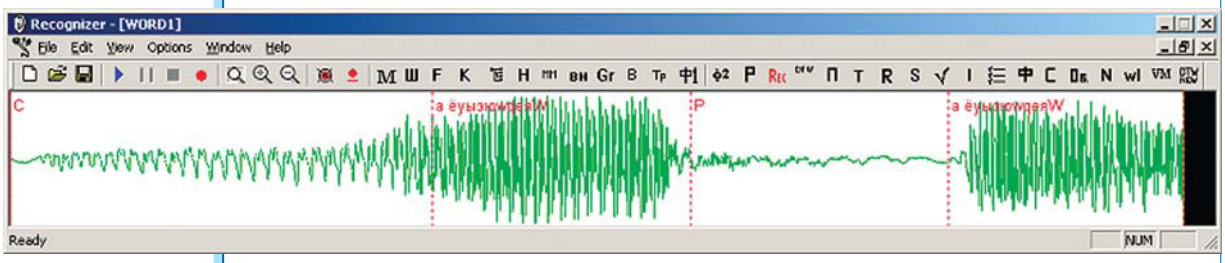

Рис. 1. Визуализация записи слова «лапа»

Добиться того, чтобы при записи речевого сигнала подобный сегмент не обрезался, если он окажется в конце, — непростая задача.

Вторая проблема конца сигнала возникает из ситуации, иллюстрируемой рис. 2. На нём приведена визуализация записи слова «она».

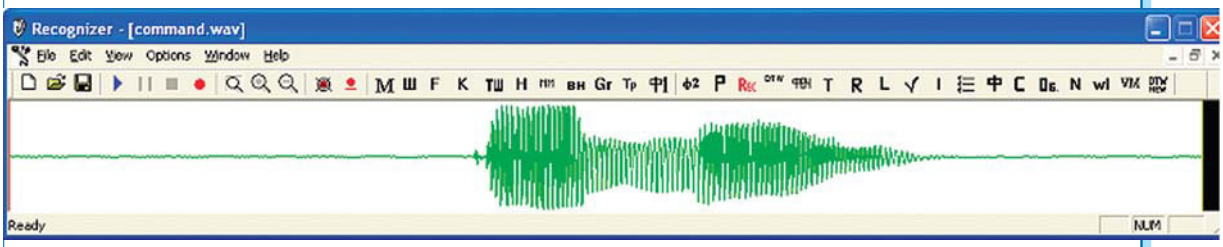

Рис. 2. Визуализация записи слова «она»

Слово произнесено через некоторое время после включения записи. Она остановлена также через некоторое время после окончания речи. Рисунок отражает тот факт, что речевой сигнал всегда затухает постепенно, при этом значительно медленнее, чем нарастает вначале. Поэтому при автоматической сегментации[1] программа может ошибочно определить в конце наличие **звонкого** согласного звука, которого на самом деле нет. Это побуждало нас долгое время отказываться от распознавания **звонких** согласных в конце слова, вплоть до исключения этих звуков из автоматически создаваемой транскрипции слов распознаваемого словаря.

В данной работе предлагается механизм решения этих проблем. Он основан на использовании величины $W(n)$, введённой в работе [1] при определении понятия вариационной меры. Напомним определение этой величины. Мы работаем с 8-битной записью при частоте дискретизации 22050 Гц, в условиях отсутствия существенного внешнего шума. Пусть x_0, x_1,... — последовательные отсчёты записанного сигнала.

Рассматривается численный аналог полной вариации с переменным верхним пределом

$$V(0)=0,\ V(n)=\sum_{i=0}^{n-1}\left|x_{i+1}-x_i\right|$$

Пусть N_1 — максимальное число такое, что $V(N_1)\le 255$. Полагаем

$W(n)=V(n)$ при $0\le n\le N_1$,

$W(n)=V(n)-256$, при $N_1+1\le n\le N_2$, где

N_2 — максимальное число такое, что $W(N_2)\le 255$ и так далее.

Таким образом, $W(n)$ — кусочно-возрастающая в широком смысле слова функция, связанная с $V(n)$ условием

$$W(n) \overset{\mathrm{mod}\,256}{=} V(n). \qquad (1)$$

При построении $W(n)$ значения $V(n)$ «сбрасываются» вниз на 256 единиц, как только они начинают превышать 255.

На рис. 3,4 приведена визуализация сигнала, отвечающего слову «сумма» и функции $W(n)$, построенной по этому сигналу.

[1] Сегментация – разбиение сигнала на участки, отвечающие отдельным звукам речи.

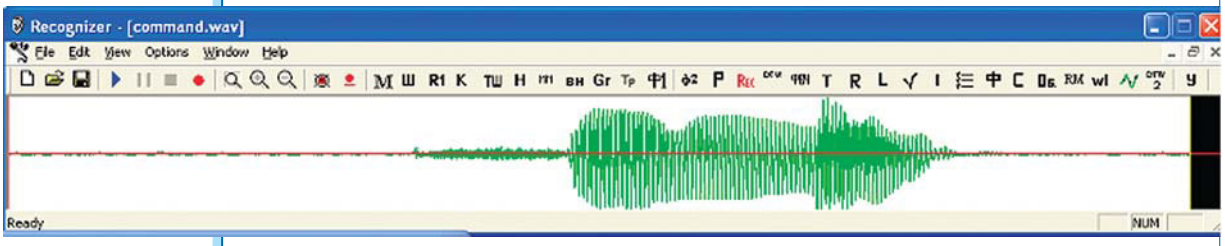

Рис. 3. Визуализация записи слова «сумма»

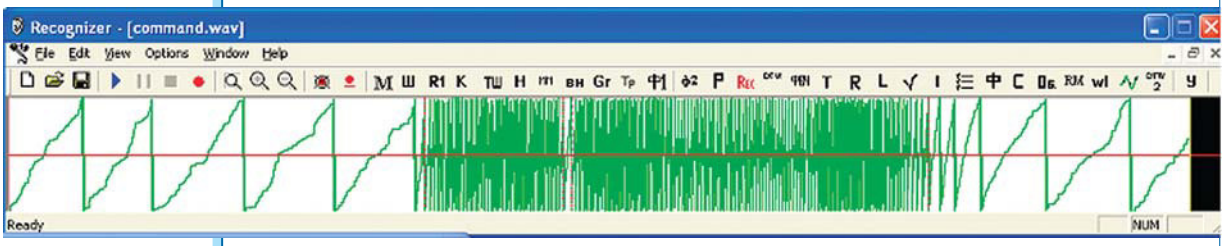

Рис. 4. Визуализация соответствующей функции $W(n)$

Участкам молчания до и после слова отвечают промежутки медленного возрастания $W(n)$ от 0 до 255 за счёт фонового шума звуковой карты.

Мы называем «сглаживанием сигнала» обработку его скользящим трёхточечным фильтром:

$$x_i = \frac{x_{i-1} + x_i + x_{i+1}}{3}.$$

Подвергнем исходный сигнал пятикратному сглаживанию. Тогда упомянутый фоновый шум **нивелируется,** и вместо сигнала на рис. 4 мы получим сигнал на рис. 5:

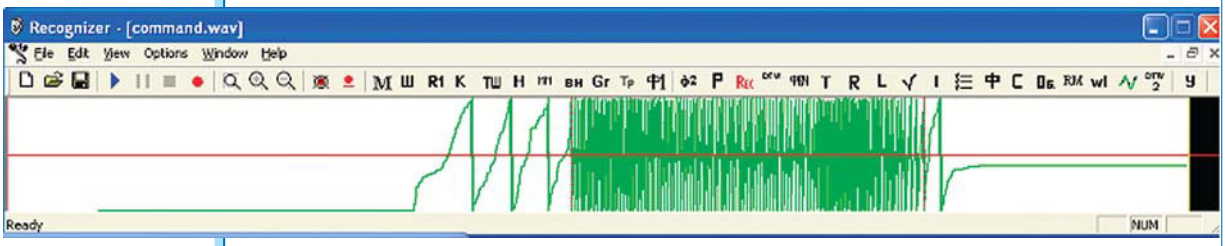

Рис. 5. Функция $W(n)$ для пятикратно сглаженного сигнала слова «сумма»

Теперь участкам молчания до и после речи соответствуют участки постоянства функции $W(n)$. Момент, когда при движении от начала записи вправо первая из этих постоянных превышается на p_1 единиц, будем считать соответствующим началу речи. Аналогично, двигаясь от конца записи влево и используя порог p_2, определяем момент окончания речи. Пороги p_1 и p_2 определяются экспериментально в зависимости от уровня шума конкретной звуковой карты и микрофона. В наших условиях мы брали их равными 10. Проведённые многочисленные эксперименты показывают, что это весьма точное определение начала и конца речи. Очевидные ошибки чрезвычайно редки. При этом, если слово заканчивается одним из глухих взрывных звуков *[К], [П], [Т],* перед концом речевого сигнала появляется достаточно длинный паузообразный участок, наличие которого позволяет

обнаруживать этот звук в конце слова. Например, визуализация записи слова «аббат» (рис. 6):

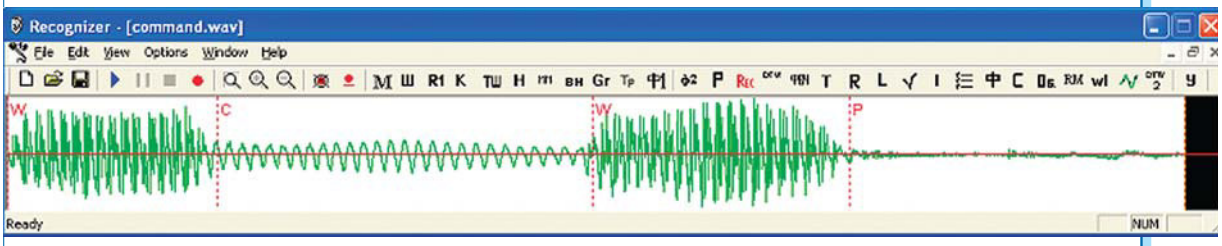

Рис. 6. Визуализация записи слова «аббат»

Всё сказанное о звуках *[К], [П], [Т]* в конце слова в равной мере относится к звукам *Ф* и *Х*, которые, **ввиду малой интенсивности также могут при распознавании восприниматься как паузообразные.** Результаты распространяются и на мягкие варианты всех этих звуков. Но следует учесть, что тогда возможно появление в конце ярко выраженной фрикативной части (как правило, у мягкого звука *[Т]*, см. рис. 7). Однако эта часть существенно короче шипящих и свистящих звуков (*[С], [Ш]* и т.д.), что позволяет не приписать сигналу этих звуков по ошибке.

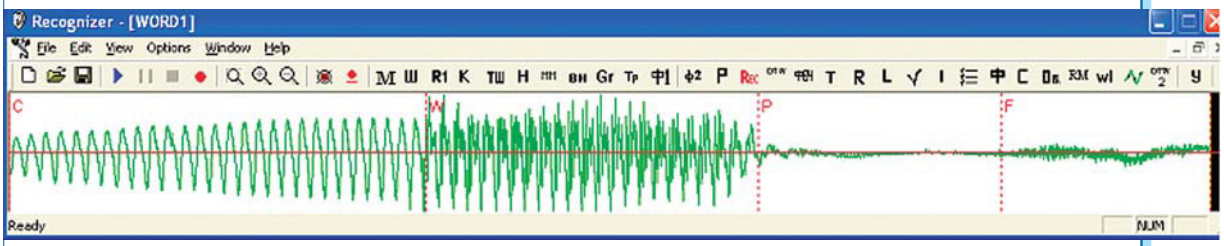

Рис. 7. Визуализация записи слова «мать»

Переходим ко второй вышеупомянутой проблеме конца сигнала. Пусть слово заканчивается голосовым звуком. Применим тот же алгоритм, заменив пятикратное сглаживание 50-кратным. Тогда заключительный участок затухания последнего голосового звука станет паузообразным, и метка конца, полученная с помощью сглаженного сигнала, передвинется влево. Этот заключительный участок, находящийся между новой меткой конца при 50-кратном сглаживании и первоначально полученной меткой конца при пятикратном сглаживании отбрасывается, и в окончательный буфер для визуализации и распознавания заносится укороченный таким образом сигнал. Как показывают эксперименты, это гарантирует при последующей сегментации от ошибочного выделения в конце слова голосового согласного звука. Алгоритм этой «последующей» сегментации предложен в работе [2], описание несколько усовершенствованного алгоритма содержится в [3]. На рис. 8 приведена визуализация записи слова «оса», где заключительный участок затухания сигнала выделен вышеописанным образом.

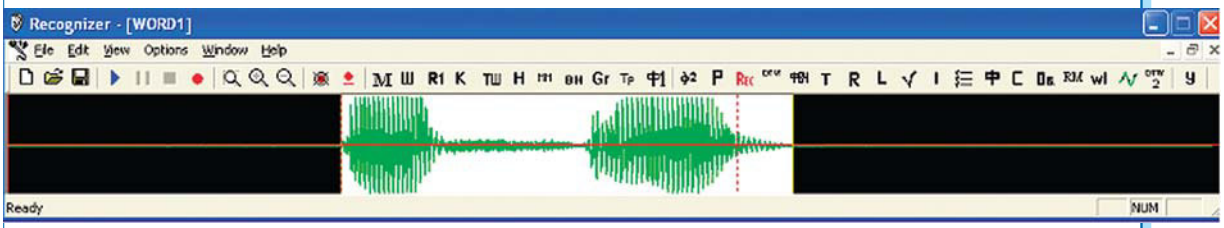

Рис. 8. Выделение заключительного участка затухания в слове «оса»

Все изложенные алгоритмы предполагают наличие сигнала, содержащего речевой отрезок и записанного с некоторым «запасом», т.е. с участками «молчания» слева и справа. Такую предварительную запись можно получить, следуя алгоритмам, изложенным в главе 1 книги [4]. При этом проверка на наличие речи с использованием квазипериодической структуры голосовых элементов позволяет осуществлять запись в автоматическом режиме. После включения пользователем начала записи компьютер будет всё время записывать один за другим слитные речевые отрезки, находясь в режиме непрерывной записи звука и ожидания речи.

Таким образом, достигается весьма эффективное решение обеих вышеописанных проблем конца сигнала.

Литература

1. *Шелепов В.Ю., Ниценко А.В., Жук А.В., Азаренко Д.С*. О распознавании фонем с помощью анализа речевого сигнала в частотной и временной областях. Приложение к распознаванию синтаксически связных фраз // Речевые технологии. М., 2008. № 2. С. 43–52.

2. *Шелепов В.Ю., Ниценко А.В*. Структурная классификация слов русского языка. Новые алгоритмы сегментации речевого сигнала и распознавания некоторых классов фонем // Искусственный интеллект. 2007. № 1. С. 213–224.

3. *Шелепов В.Ю., Ниценко А.В., Жук А.В*. Построение системы голосового управлении компьютером на примере задачи набора математических формул // Искусственный интеллект. 2010. № 3. С. 259–267.

4. *Шелепов В.Ю*. Лекции о распознавании речи. Донецк: ІПШІ «Наука і освіта», 2009. 196 с.

Сведения об авторах

Шелепов Владислав Юрьевич —
доктор физико-математических наук, профессор, автор ряда работ в ведущих математических журналах СССР. С 1993 г. возглавлял отдел распознавания речевых образов Института проблем искусственного интеллекта НАН и МОН Украины. В настоящее время — заведующий кафедрой систем искусственного интеллекта Института информатики и искусственного интеллекта Донецкого национального технического университета.

Ниценко Артём Владимирович —
профессиональный программист, специализирующийся в области распознавания речи, автор многочисленных экспериментальных и прикладных программ.