УДК 004.934.1'1

Е.Е. Федоров, В.Ю. Шелепов
Институт проблем искусственного интеллекта, г. Донецк

Автоматическое определение начала и конца записи речи

В статье описываются разработанные авторами алгоритмы определения моментов, когда включенный распознаватель начинает и заканчивает «слышать» слитный речевой фрагмент. Они опробованы в отделе распознавания речи Института проблем искусственного интеллекта и показали высокую надежность и эффективность.

Перед началом выполнения алгоритма записывается сигнал длиной в 300 отсчетов (8-битная запись, частота дискретизации – 22050 Гц) и вычисляется величина средней линии

$$AverageLine = \frac{\sum_{i=0}^{299} x_i}{300},$$

где x_i – значение i-го отсчета сигнала.

Вычисляется также максимальное отклонение для шума

$$MaxDeviationNoise := \max_{0 \le i \le 299} \left| x_i - AverageLine \right|.$$

Алгоритм. Определение начала речи в сигнале с учетом квазипериода

Шаг 1. $NumPeriod := 0,\ NumPeriod1 := 0,\ MIN := 60,\ MAX := 200,\ n_0 := 0.$

Шаг 2. Получение очередной порции звуковых данных длиной в 300 отсчетов, которые помещаются в конец буфера x. Вычисляется величина

$$L_k := \sum_{i=0}^{k-1} \left| x_{n_0+i} - x_{n_0+i+k} \right|, \qquad (1)$$

где n_0 – номер отсчета, с которого в буфере начался текущий квазипериод,

$MIN \le k \le MAX$.

Определяется k_0, при котором L_k минимальна. По определению, k_0 представляет собой длину квазипериода.

Если $k_0 \ge MIN+20,$

то вычисляется среднее отклонение для речи

$$CurMeanDeviationSpeech := \frac{\sum_{i=0}^{k_0-1} \left| x_{n_0+i} - AverageLine \right|}{k_0}$$

Если CurMeanDeviationSpeech > MaxDeviationNoise, (2)
то *NumPeriod* : = *NumPeriod* +1,
иначе *NumPeriod* : = 0, переход на шаг 3.
Если количество квазипериодов *NumPeriod* > 3, (3)
то переход на шаг 4,
иначе $n_0 := n_0 + k_0$, переход на шаг 2.
Если $k_0 <$ *MIN+20*,
то находится количество строгих минимумов на текущем квазипериоде и заносится в *NumMin.*
Если *NumMin* ≥ 3, (4)
то *NumPeriod1* : = *NumPeriod1* +1,
иначе *NumPeriod1* : = 0, переход на шаг 3.
Если NumPeriod1 > 5, (5)
то переход на шаг 4,
иначе $n_0 := n_0 + k_0$, переход на шаг 2.

Шаг 3. Удаление из буфера первых $n_0 + k_0$ элементов, $n_0 : = 0$, переход на шаг 2.

Шаг 4. Завершение алгоритма

Условие $k_0 \geq$ *MIN+20* используется для учета случая, когда речевой фрагмент начинается с голосового звука, а условие $k_0 <$ *MIN+20* – для учета случая, когда фрагмент начинается с шипящей (шипящие не содержат квазипериода, поэтому минимизация функции (1) происходит за счет минимизации числа слагаемых). Условие (2) используется для учета шума.

Выбор конкретных значений MIN = *60,* MAX = *200 ориентирован на участвовавших в экспериментах дикторов (квазипериод от 80 до 180 отсчетов). В общем случае эти величины выбираются автоматически после определения квазипериода диктора во время предварительной настройки системы перед запуском распознавателя.*

На рис. 1 представлена схема определения начала речи. Переменная MDN соответствует переменной *MaxDeviationNoise*, переменная CMDS соответствует переменной *CurMeanDeviationSpeech*.

После записи сигнала осуществляется его дополнительная проверка на шум. Если число найденных с помощью минимизации величины (1) квазипериодов, идущих друг за другом и имеющих длину $k_0 \geq$ *MIN+20*, превышает 5, то считается, что записанный сигнал содержит речь и подлежит распознаванию [3]. В противном случае записанный фрагмент считается шумом и распознаванию не подлежит.

В основе алгоритма определения конца речи, то есть момента, когда вслед за речевым фрагментом начинается пауза, лежит тот факт, что при 8-битной записи соответствующий отрезок сигнала имеет большие участки с постоянной амплитудой. Следовательно, для него отношение общего числа отсчетов к числу нестрогих минимумов достаточно мало.

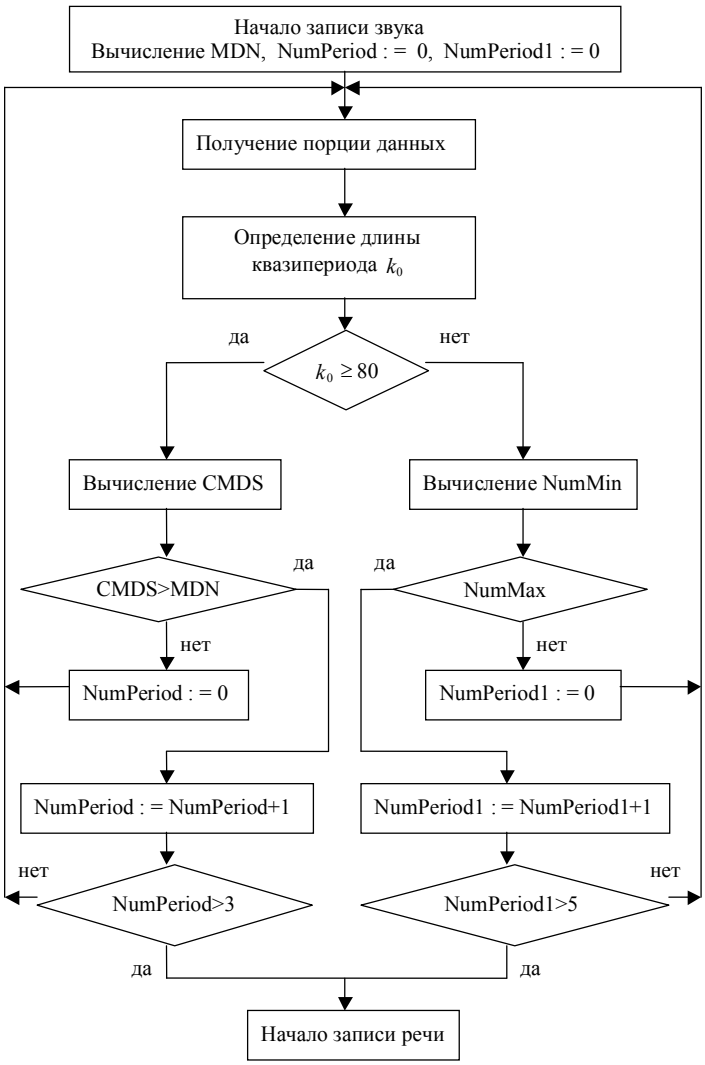

Рис. 1. Блок-схема определения начала речи

АЛГОРИТМ. Определение конца речи в сигнале

Шаг 0. *NumPeriod2* : = 0

Шаг 1. В очередной порции звуковых данных определяется количество нестрогих минимумов *NumMin1*

Шаг 2.

$$\text{Если} \quad \frac{300}{NumMin1} < 2, \tag{6}$$

то *NumPeriod2* : = *NumPeriod2* + 1,

иначе *NumPeriod2* : = 0.

Шаг 3.

$$\text{Если} \quad NumPeriod2 > 15, \tag{7}$$

то переход на шаг 4,

иначе переход на шаг 1.

Шаг 4. Завершение алгоритма.

Константы в формулах (3) – (7) определены эксперементально. Они частично зависят от используемого микрофона и звуковой карты и при реализации алгоритмов в конкретных условиях могут потребовать корректировки.

Сравнение с другими методами

В литературе описаны следующие методы определения речи в сигнале. Применяемый в мобильных телефонах детектор активности речи (VAD) [4], основанный на различии спектральных характеристик речи и шума, несмотря на надежность является достаточно трудоемким, а использование фильтрации оказывает негативное влияние на записанный сигнал [5], что сказывается на надежности распознавания. Детектор речи КБ Спецвузавтоматики [6], основанный на спектре мощности, имеет существенный недостаток – превышение верхнего порога может быть связано со случайной высокоамплитудной помехой. Детектор речи Рабинера и Сэмбура [7], использующий энергию сигнала и число переходов через нуль, также не учитывает случайную высокоамплитудную помеху. Что касается коммерческих систем типа Dragon Dictate и ViaVoice, то они сильно зависят от типа используемого микрофона.

Разработанный в данной статье метод не оказывает влияния на записанный сигнал и достаточно надежно определяет границы речи, учитывая случайную высокоамплитудную помеху.

Литература

1. Федоров Е.Е. Система голосового управления // Труды конф. «Информационные технологии в науке, образовании, телекоммуникации, бизнесе». – Запорожье. – 2001. – С. 104-106.
2. Федоров Е.Е., Шелепов В.Ю. Защита речевых распознавателей от шума и посторонней речи // Искусственный интеллект. – 2001. – № 3. – С. 584-587.
3. Златоустова Л.В. Фонетические единицы русской речи. – М.: Моск. ун-т., 1981. – 108 с.
4. Freeman D., Sonthcott C., Boyd I. A Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service // IEE Colloquium «Digitized Speech Communication via Mobile Radio». – London. – 1988. – P. 61-65.
5. Секунов Н.Ю. Обработка звука на PC. – СПб.: БХВ – Санкт-Петербург, 2001. – 1248 с.
6. Аграновский А.В., Зулкарнеев М.Ю., Леднов Д.А., Репалов С.А. Организация иерархической модели распознавания слитной речи // Искусственный интеллект. – 2001. – № 3. – С. 17-22.
7. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. – М.: Радио и связь, 1981. – 495 с.

In the article in described the algorithms of specification of moments when the switched on recognizer begins and finishes «hear» the streams of speech. They mere tested in the department of speech recognition of Institute of artificial intelligence (Donetsk) and showed their high efficiency.

Статья поступила в редакцию 01.07.02.