6К

УДК 004.93

А.А. Карпов
Санкт-Петербургский институт информатики и автоматизации РАН, Россия,
karpov@iias.spb.su

Робастный метод определения границ речи на основе спектральной энтропии

Для распознавания речи, как изолированной, так и слитной, необходимо предварительно определить ее границы в контексте окружающих бесполезных звуковых сигналов. Сложность определения границ речи связана с особенностями произношения конкретного диктора, наличием в речевом сигнале различных видов посторонних шумов, а также звуковых артефактов процесса артикуляции (придыхание, чмокание и т.п.). Традиционные методы, основанные на вычислении кратковременной энергии сигнала, спектральной энергии или количестве нуль-пересечений неустойчиво работают в условиях, когда появляются шумы с динамическим спектром или относительно сильные стационарные шумы. В статье рассматривается метод определения границ речи, основанный на вычислении энтропии (как меры неопределенности или беспорядка в некотором распределении) спектра сигнала. Для робастного выделения речи используется свойство отличия значений энтропии для речевых сегментов и для фоновых шумов. Отличительная черта данного подхода состоит в том, что этот показатель является малочувствительным к изменениям амплитуды сигнала и, следовательно, позволяет более робастно и точно определять границы речи. Экспериментальные результаты по применению разработанного метода показали, что речевые фрагменты успешно выделяются из звуковых сигналов, содержащих различные виды сильных шумов (белый, коричневый, розовый, узкополосный шум и т.д.) и звуковых артефактов. Кроме того, разработанный метод имеет приемлемую вычислительную сложность, что позволяет его эффективно использовать в системах распознавания речи реального времени.

Введение

Традиционно в системах распознавания речи для определения границ речи используются методы (например, Voice Activity Detector), основанные на вычислении кратковременной энергии сигнала или спектральной энергии. Кроме того, дополнительно применяются методы, использующие количество нуль-пересечений сигнала и информацию о длительности речевых фрагментов. Однако все эти алгоритмы становятся менее надежными в условиях нестационарного шума, а также при возникновении различных звуковых артефактов (придыхание, чмокание и т.п.). Также существуют алгоритмы, основанные на адаптивных пороговых значениях, но при возникновении звуковых артефактов, а также относительно высоком уровне шума или незначительном уровне полезного сигнала они также становятся неустойчивыми. Поэтому была поставлена задача разработать эффективный метод для определения границ речи, который позволил бы устойчиво выделять речь при наличии нестационарного шума.

При разработке к методу определения границ речи предъявляются следующие основные требования:

- обеспечение минимальной вероятности ложного срабатывания при воздействии только шума с высоким уровнем;
- высокая вероятность правильного выделения речи даже в условиях сильного шума;
- высокое быстродействие для исключения задержек включения и выключения распознавателя речи.

1 Математическая основа метода

Разработанный нами метод основан на вычислении энтропии (как меры неопределенности или беспорядка в некотором распределении [1]) спектра сигнала. Для определения границ речи используется свойство отличия значений энтропии для речевых сегментов и для фонового шума. Отличительная черта данного подхода состоит в том, что этот показатель является мало чувствительным к изменениям амплитуды сигнала. Впервые энтропию спектра было предложено использовать для данной задачи всего несколько лет назад [2], [3]. Разработанный нами метод является развитием данных идей и добавляет несколько новых уровней при анализе звукового сигнала. Рис. 1 иллюстрирует блок-схему алгоритма метода определения границ речи на основе спектральной энтропии.

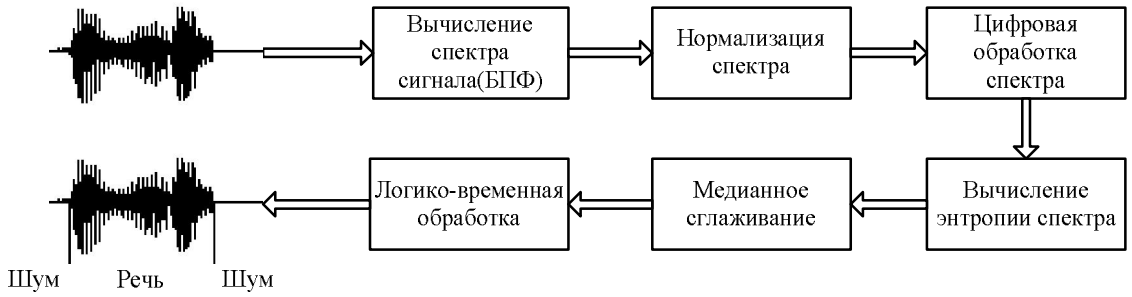

Рисунок 1 – Алгоритм определения границ речи на основе энтропии спектра сигнала

Работа алгоритма осуществляется следующим образом. Поступающий с микрофона сигнал оцифровывается с частотой дискретизации 16 КГц и делится на короткие сегменты по 16 мс, содержащие по 256 цифровых отсчетов. При этом перекрытие соседних сегментов составляет 70 отсчетов (чуть более 25 %).

Далее, используя алгоритм быстрого преобразования Фурье (БПФ), вычисляется кратковременный спектр сегмента сигнала. Затем производится нормализация вычисленного спектра по всем частотным компонентам:

$$p_i = \frac{s(f_i)}{\sum_{k=1}^{N} s(f_k)}, \quad \text{i} = 1\ldots\text{N}$$

где $s(f_i)$ – спектральная энергия для спектральной компоненты f_i, p_i – соответствующая плотность вероятности, а N – количество спектральных компонент в БПФ. Полученная функция представляет собой функцию плотности вероятности спектра. Количество используемых спектральных компонент может выбираться от нескольких десятков до нескольких сотен. Здесь важно найти компромисс между необходимой чувствительностью и вычислительной нагрузкой. В нашей модели используется 256 спектральных компонент.

Для того чтобы уже на этом этапе анализа отбросить некоторые виды шумов, необходимо ввести некоторые ограничения, а именно: полоса частот ограничивается значениями 200 – 8000 Гц, т.е.

$$s(f_i) = 0, \text{ if } f_i < 200 \text{ Гц}.$$

Эта полоса частот охватывает практически все частотные компоненты, присутствующие в речевом сигнале человека. Такое ограничение позволяет исклю-

чить воздействие как очень низкочастотных, так и высокочастотных шумов (например, внутренних шумов звуковой карты или микрофона).

Возможные значения плотности вероятности ограничиваются как сверху, так и снизу, что позволяет исключить шумы, сосредоточенные в узкой частной области, а также шумы, имеющие приблизительно одинаковое распределение частотных компонент по всему спектру (например, белый шум).

$$p_i = 0, \text{ if } p_i < \delta_2 \text{ or } p_i > \delta_1,$$

где δ_1 и δ_2 – верхняя и нижняя границы плотности вероятности, соответственно. В нашей модели значения δ_1 и δ_2 равны 0,3 и 0,01 соответственно.

Дополнительно могут использоваться различные методы очистки сигнала от шума (например, адаптивный фильтр Кальмана или методы спектрального вычитания) [4].

На следующем этапе производится вычисление спектральной энтропии полученного нормированного спектра по следующей формуле [3]:

$$H = -\sum_{k=1}^{N} p_k \log p_k$$

На следующем шаге анализа применяется медианное сглаживание последовательности полученных значений спектральной энтропии ξ. В отличие от многих других методов сглаживания (например, метода скользящих средних), данный метод является значительно более устойчивым к отдельным выбросам и случайным искажениям данных. В основе метода лежит вычисление скользящей медианы. Для того чтобы найти значение скользящей медианы в точке *t*, вычисляется медиана значений ряда во временном интервале *[t-q, t+q]*. Медиана ряда во временном интервале определяется как центральный член последовательности значений ряда, входящих в этот временной интервал, упорядоченной по возрастанию.

В ходе экспериментов наилучшие результаты показал метод медианного сглаживания в окне размером 5. Однако, если момент времени *t* отстоит от начала или конца ряда менее чем на *q* точек, вычисление становится невозможным. Поэтому здесь для устранения таких краевых эффектов вычисляется значение скользящей медианы для меньшего, но максимально возможного окна.

Далее для некоторых задач, в которых вид шума и его спектр мало меняются с течением времени, может оказаться эффективным дополнительное вычисление энтропии спектра для короткого участка звукового сигнала, содержащего только акустический фоновый шум без включения речевых или иных звуковых фрагментов, и вычитание энтропии для шума из полученных значений энтропии для анализируемого сигнала.

На последнем этапе применяется логико-временная обработка, учитывающая допустимые на практике длительности речевых и неречевых фрагментов. Сначала вычисляется адаптивный порог, который служит для выделения краевых точек (начала и конца) гипотезы фрагмента речи:

$$\gamma = \left(\frac{\max(\xi) - \min(\xi)}{2} + \min(\xi) \right) * \mu,$$

где μ – коэффициент, который подбирается экспериментальным путем. В нашей модели данный коэффициент принимает значения от 0,8 до 1,1 в зависимости от зашумленности сигнала. На порог γ, в свою очередь, также накладывается ограничение на минимальное значение, достоверно определяющее речевой фрагмент. В нашей системе минимальное допустимое значение порога равно 1,6 (при возможных на

практике значениях энтропии спектра от 0 до 3 единиц). На основе этого порога выбираются акустические сегменты анализируемого сигнала, которые принадлежат к речи человека. После этого производится логико-временная обработка выделенных участков сигнала. Эта обработка необходима, так как во многих случаях из-за возникновения различных звуковых артефактов безречевые участки сигнала ошибочно принимаются за речь, и наоборот, некоторые участки, содержащие речь, отбрасываются из-за специфических акустических характеристик. При логико-временной обработке применяются два основных показателя (рис. 2):

- минимальные длительности (s_i и s_j) выделенных фрагментов, содержащих речь (на рис. 2 – области a_ib_i и a_jb_j);
- максимальная длительность (n_i) безречевого участка (область b_ia_j) между двумя соседними выделенными речевыми фрагментами.

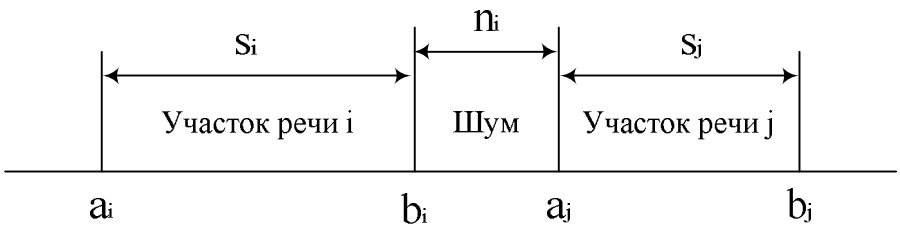

Рисунок 2 – Логико-временная обработка

Учитывая, что человек не может производить очень короткие речевые фрагменты, а также то, что в речи всегда присутствуют определенные паузы (например, смычки перед взрывными согласными), мы экспериментальным путем определили пороговые значения для минимальной длительности речевого участка и максимальной длительности безречевого участка. Если речевые участки (области a_ib_i и a_jb_j) в некотором сигнале имеют длительности не менее 15 сегментов, а безречевой участок между ними (область b_ia_j) – не более 20 сегментов, то все данные фрагменты участки объединяются в один речевой фрагмент (отрезок a_ib_j), который будет являться результатом работы алгоритма.

2 Экспериментальные результаты

В этом разделе представляются некоторые результаты тестирования предложенного метода для выделения границ речи. Был проведен ряд опытов по отделению раздельно произнесенных слов, а также слитно произнесенных фраз от фонового сигнала, в который были искусственно добавлены различные виды шумов, сгенерированные автоматически при помощи программы CoolEdit Pro. Для проверки работоспособности метода использовались следующие виды шумов:

1) шум с узкой полосой частот (2700 – 3300 Гц). Данный шум можно приближенно считать монотонным сигналом с частотой 3000 Гц;

2) белый шум. Данный шум имеет спектр с приблизительно постоянной спектральной плотностью в полосе частот от 0 до 8000 Гц;

3) коричневый шум. Спектральная плотность уменьшается на 6 дБ с каждой последующей октавой (т.е. спектральная плотность обратно пропорциональна квадрату частоты);

4) розовый шум. Спектр такого шума имеет спектральную плотность, уменьшающуюся на 3 дБ с каждой последующей октавой (спектральная плотность обратно пропорциональна частоте);

5) усиленный акустический фон, записанный в помещении, в котором производились эксперименты.

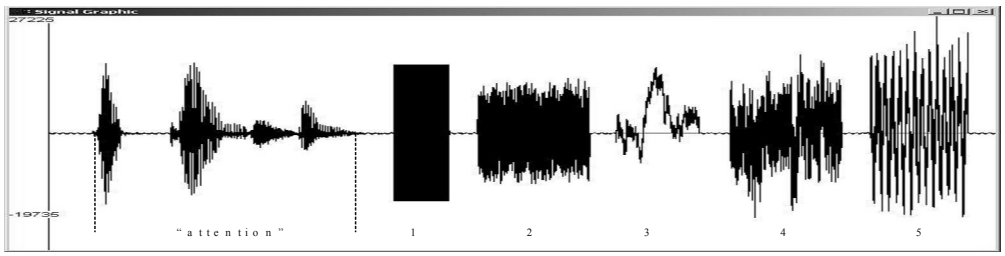

Рисунок 3 – Тестовый сигнал

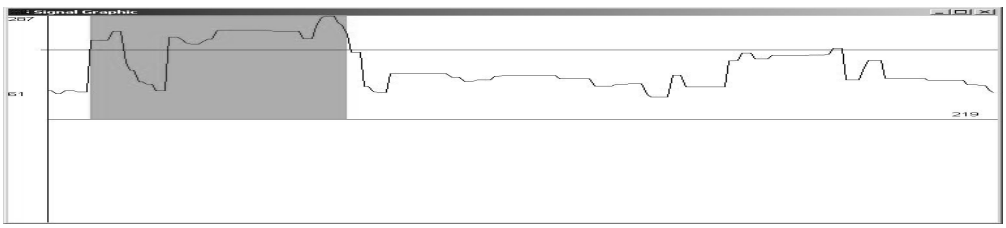

Рисунок 4 – Результат выделения границ речи

Рис. 3 и 4 иллюстрируют результаты по выделению произношения слова «attention» из сигнала, в котором присутствуют все вышеперечисленные виды шумов с амплитудами большими или равными произнесенной речи. Очевидно, что речь из данного сигнала при помощи разработанного алгоритма была выделена абсолютно точно. Также можно заметить, что со всеми видами шумов алгоритм справился, не перепутав их с речью.

На рис. 5 и 6 представлены результаты по выделению произношения слова «attention» из сигнала, который получился смешиванием исходного звукового сигнала с псевдослучайным белым шумом большой амплитуды.

В данном эксперименте отношение сигнал/шум (SNR) равнялось около 3 ДБ. Из рисунков видно, что алгоритм прошел это испытание также успешно.

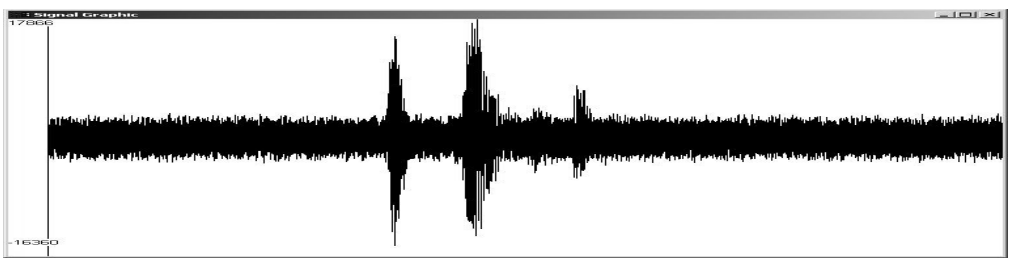

Рисунок 5 – Тестовый сигнал

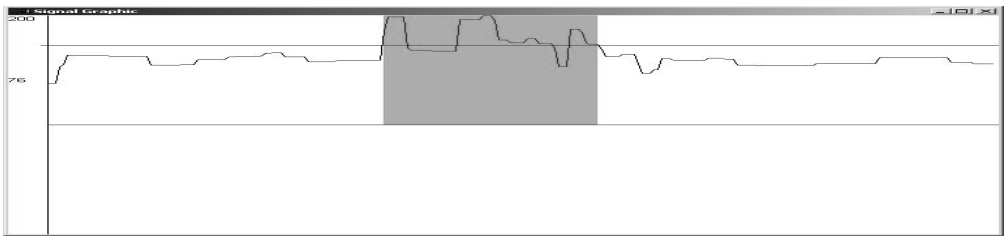

Рисунок 6 – Результат выделения границ речи

Кроме того, мерой эффективности алгоритма может служить величина ошибки выделения границ речи [5], которая складывается из двух показателей:

– вероятности $P_{ложн.идент.}$ ложной идентификации, т.е. определения факта наличия речевого сообщения в момент времени, когда речевого сообщения на самом деле не было;
– вероятности $P_{усеч.сегм.}$ усечения речевого сегмента, т.е. отсутствия сигнала о наличии речевого сообщения в момент времени, когда речь в сигнале присутствует.

Таким образом, ошибка выделения границ речи равна: $P_{ош.гр.} = P_{ложн.идент.} + P_{усеч.сегм.}$

В ходе экспериментов были получены следующие результаты (табл. 1) по автоматическому выделению раздельно произносимых английских слов из сигнала при наличии в нем различных видов шумов, сопоставимых по уровню с полезным сигналом.

Табл. 1 отражает высокую эффективность алгоритма при воздействии всех вышеперечисленных видов нестационарных шумов (наихудший случай – это присутствие розового шума, который по своей форме напоминает настоящий речевой сигнал). Ошибка, выражающаяся в усечении речевого участка, остается практически постоянной для всех экспериментов. Это объясняется периодическим усечением смычки перед взрывными согласными («t», «d», «p») в начале слове. Данный участок часто воспринимается алгоритмом как тишина. Однако данная проблема успешно решается на дальнейших уровнях обработки при распознавании речи.

Таблица 1 – Ошибка выделения границ при различных видах шумов

Вид шума		Вероятность ложной идентификации, %	Вероятность усечения речи, %	Ошибка выделения речи, %
1	Узкополосный шум	0	2	2
2	Белый шум	1	2	3
3	Коричневый шум	3	2	5
4	Розовый шум	15	3	18
5	Усиленный акустический фон	2	2	4

Заключение

Основные трудности задачи определения границ речи обусловлены изменчивостью произнесения и разнообразием используемого словаря, наличием специфических пауз перед смычками внутри слов, воздействием нестационарного

шума. Для решения этой задачи был разработан метод, основанный на вычислении энтропии кратковременного спектра звукового сигнала. Экспериментальная проверка показала, что речевые фрагменты успешно выделяются из звуковых сигналов, содержащих различные типы сильных фоновых шумов и звуковых артефактов. Кроме того, разработанный метод имеет достаточно высокую производительность, что позволяет его эффективно использовать в системах распознавания речи реального времени.

Литература

1. Jitendra Ajmera, Iain McCowan, Herve Bourlard. Speech/music segmentation using entropy and dynamism features in a HMM classification framework // Speech Communication. – 2003. – Vol. 40. – P. 351-363.
2. Khurram Waheed, Kim Weaver and Fathi M. Salam. A robust algorithm for detecting speech segments using an entropy contrast // Proc. 45th IEEE International Midwest Symposium on Circuits and Systems MWSCAS'2002. – Oklahoma (USA). – 2002.
3. Shen J.-L., Hung J.-W., Lee L.-S. Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments // Proc. Int. Conf. on Spoken Lang. Processing ICSLP'98. – Sydney (Australia). – 1998.
4. Masakiyo Fujimoto, Yasuo Ariki. Evaluation of noisy speech recognition based on noise reduction and acoustic model adaptation on the AURORA2 tasks // Proc. Int. Conf. on Spoken Lang. Processing ICSLP'2002. – Denver (USA). – 2002.
5. Мазуренко И.Л. Многоканальная система распознавания речи // Сборник трудов VI всероссийской конференции «Нейрокомпьютеры и их применение». – Москва. – 2000.

The Robust Method for Speech Endpoint Detection Based on Spectral Entropy

To recognize both isolated and continuous speech it is required to detect the boundaries of speech in context of surrounding useless sound signals. The complexity of speech endpoint detection is connected with the peculiarities of the pronunciation of concrete speaker, presence of diverse noises in speech signal and sound artifacts (aspiration, lip smacks, etc) in articulation process. The traditional methods for speech detection based on calculation of short-time signal energy, spectral energy or amount of zero-crossings work unrobustly when there are any dynamic noises or stationary noises with high amplitude. In this paper the method for speech detection based on calculation of entropy (as the measure of uncertainty or disorder in a given distribution) of signal spectrum is considered. The distinction between entropy for speech segments and entropy for background noises is used for robust speech endpoint detection. The important feature of this method is that such method is less sensitive to the variations of the signal amplitude and, hence, allows detecting speech boundaries more robustly and exactly than traditional ones. The experimental results of usage of the developed method have shown that speech fragments are successfully selected in sound signals, which contain diverse kinds of intense noises and sound artifacts. Moreover, this method has sufficiently high speed of processing and can be used in real-time speech recognition systems.

Статья поступила в редакцию 29.06.2004.