КРАТКИЕ СООБЩЕНИЯ

УДК 534.4

ОБНАРУЖЕНИЕ РЕЧИ В СИГНАЛАХ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ

Д.А. БОРИСЕВИЧ, Г.В. ДАВЫДОВ, В.А. ПОПОВ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 18 апреля 2013

Рассмотрены информационные параметры речевых сигналов, позволяющих выделить участки обрабатываемой информации с речевыми сигналами в реальном времени. Выполнено моделирование в программном пакете MatLab по обнаружению речи в сигналах в реальном времени.

Ключевые слова: речевые сигналы, информационные параметры, мощность сигнала, количество переходов через ноль, стационарность сигнала, кепстр.

Введение

Обнаружение речи в сигналах и выделение участков с речевой информацией является весьма важным аспектом при обработке речи и находит широкое применение в устройствах защиты речевой информации, в кодеках речи, в речевых интерфейсах управлений и других областях, связанных с необходимостью выделения во времени участков в сигналах, несущих речевую информацию. Обнаружить речь в сигналах с наибольшей вероятностью можно, используя информационные признаки речевых сигналов.

Информационный параметр – мощность сигнала

Одним из информационных параметров для детектирования речи является превышение порогового уровня мощности сигнала. При использовании данного информационного параметра задается порог, обычно на уровне, устанавливаемом экспериментально, в зависимости от уровня шума в канале или акустического шума в помещении, когда необходимо определить моменты времени появления речевых сигналов в данном помещении. Относительно этого порога и происходит разделение сигналов. Также при использовании данного информационного параметра порог может задаваться как уровень мощности фонового шума. В таком случае уровень мощности шума измеряется в те отрезки времени, когда присутствует только шум. Зачастую уровень мощности шума не постоянен во времени, поэтому иногда используются алгоритмы измерения уровня мощности шума в реальном времени, которые позволяют постоянно изменять порог в зависимости от фонового шума [1, 2].

При использовании данного информационного параметра не рассматриваются какие-либо специфические характеристики сигнала присущие только речевым сигналам, поэтому различные импульсные помехи, хлопки, удары будут детектироваться как речь. Данный информационный параметр не является оптимальным, но зачастую используется для предварительного детектирования в различных системах и устройствах детектирования речи. Также он не позволяет отделить во времени участки с речевой информацией от участков с другими видами сигналов. Используя данный информационный параметр можно сократить объем вычислений путем исключения из обработки по другим информационным параметрам участков, где мощность сигнала меньше порогового уровня.

Для повышения разрешающей способности устройств выделения участков с речевой информацией (детекторов речи) и уменьшения ложных срабатываний в них используются информационные параметры, которые базируются на особенностях речевых сигналов.

Информационный параметр – количество пересечений с нулем

К информационным параметрам, которые описывают особенности речевого сигнала во временной области, можно отнести информационный параметр количества пересечений с нулем. Количество пересечений с нулем [2] – характеристика, которая показывает, сколько раз на временном интервале сигнал пересекает нулевой уровень. Если два последовательных отсчета имеют различные знаки, то произошел переход сигнала через нуль. При использовании информационного параметра количества пересечений с нулем, анализируемый цифровой сигнал $x(n)$ предварительно фильтруется низкочастотным фильтром с передаточной функцией $H(z)$: $H(z) = \frac{1-a}{1-az^{-1}}$, где a = 63/64, $z = Ae^{j\varphi}$ – комплексное число.

Затем сигнал разбивается на кадры длительностью 64 мс. При частоте дискретизации 8 кГц в кадре получается 512 отсчета. Для каждого кадра рассчитывается количество пересечений с нулем zc по следующей формуле: $zc = \sum_{n=0}^{N-1} |\operatorname{sgn}[x(n)] - \operatorname{sgn}[x(n-1)]|$, где $x(n)$ – отсчеты сигнала в кадре, N – количество отсчетов в кадре, $\operatorname{sgn}[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases}$.

Если $zc > 24$, то принимается решение, что кадр содержит речь.

Стоит отметить, что различные помехи, такие как хлопки и удары по столу, также могут классифицироваться как речь при использовании данного информационного параметра. Также при использовании данного информационного параметра музыка будет распознаваться как речь, так как во временной области для многих музыкальных инструментов число пересечений с нулем больше 24. Как речь при использовании данного информационного параметра могут классифицироваться и сигналы всплескового характера с присутствием шумов. После прохождения всплескового сигнала в течении паузы шумы могут давать число переходов через нуль больше 24.

Информационный параметр – динамика изменения мощности сигнала

Для речи характерно постоянное изменение мощности сигнала. Так при возрастании и затухании вокализованных участков речи уровень мощности стремительно увеличивается или уменьшается. Эти изменения уровня мощности позволяют рассчитать информационный параметр динамики мощности [2]. Для расчета информационного параметра динамики мощности $sl(i)$ цифровой сигнал $x(n)$ кодируется по μ-закону (алгоритм аналогового сжатия для модификации динамического диапазона аналогового речевого сигнала до преобразования в цифровой вид). Затем цифровой сигнал $y(n)$ фильтруется низкочастотным фильтром с передаточной функцией $H(z)$: $H(z) = \frac{1-a}{1-az^{-1}}$, где a = 63/64, $z = Ae^{j\varphi}$ – комплексное число.

Затем сигнал разбивается на кадры длительностью 32 мс. При частоте дискретизации 8 кГц в кадре получается 256 отсчета. Для двух соседних кадров рассчитывается информационный параметр динамики мощности $sl(i)$ по следующей формуле [2]: $sl(i) = \sum_{i=0}^{N-1} |y(i) - y(i + 8 \cdot 32)|$, где $y(i)$ – отсчеты сигнала в кадре, N – количество отсчетов в кадре.

Если $sl(i) > 10$, то принимается решение, что кадр содержит речь. Это аналогично тому, что если огибающая речевого сигнала (или модулирующий сигнал) меньше 10 Гц, но больше 4 Гц, то принимается решение, что сигнал является речевым.

Детектирование в спектральной области

К информационному параметру, который рассматривает особенности речевого сигнала в спектральной области, относится стационарность спектра [3]. Цифровой сигнал разбивается на кадры по 512 отсчетов, при частоте дискретизации 22050 Гц длительность одного кадра составляет примерно 23 мс. Анализ на наличие речи проводится для K = 6 кадров. Для каждого кадра с помощью дискретного преобразования Фурье рассчитывается спектр: $C(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n) e^{-i\frac{2\pi kn}{N}}$, где $x(n)$ – отсчеты сигнала в кадре, N – количество отсчетов в кадре.

Затем рассчитывается амплитудный спектр по формуле: $A = \sqrt{\mathrm{Re}^2(C(k)) + \mathrm{Im}^2(C(k))}$,

где $\mathrm{Re}(C(k))$ и $\mathrm{Im}(C(k))$ – соответственно действительная и мнимая части коэффициента Фурье.

Далее осуществляется усреднение полученного спектра по восьми октавным полосам, после чего рассчитывается решающая функция по формуле $D = \frac{1}{8(K+1)}\sum_{k=0}^{K}\sum_{i=1}^{8} | A_{t+k,i} - A_{t+k-1,i} |$,

где $A_{t+k,i}$ – усредненное значение спектра для i-й октавной полосы, t – номер кадра, K – количество кадров.

Если значение решающей функции D не равно нулю, то принимается решение, что в буфере из 6 кадров присутствует речь.

Информационный параметр стационарности

Информационный параметр стационарности на базе кепстрального метода детектирует речь путем оценки корреляционных свойств кепстрального вектора [4]. Допущение не стационарности воспроизводится в кепстральных коэффициентах. Цифровой сигнал делится на кадры $x(n)$ длиной N. Затем для каждого кадра рассчитывается спектр $s(k)$. И для каждого кадра на основе спектра рассчитываются кепстральные коэффициенты. Кепстр определяется последовательностью коэффициентов разложения функции в степенной ряд. В данном случаи кепстральные коэффициенты определяются следующим выражением: $c(q) = \frac{1}{2\pi}\sum_{k=0}^{N-1} \ln(\mathrm{s}(k))^2 e^{ikq}$,

где N – количество отсчетов в кадре.

Затем рассчитывается усредненное евклидово расстояние для кепстральных коэффициентов по следующей формуле:

$$R = \frac{\sqrt{\sum_{q=0}^{N-1} c^2(q)}}{N}.$$

Большое значение усредненного евклидового расстояния (R > 50) является хорошим признаком не стационарности сигнала (т.е. речи). В то время как маленькое усредненное евклидово расстояние (R < 10) является хорошим индикатором стационарности сигнала (т.е. не речь). Кепстральный метод хорошо определяет большинство сигналов, но не определяет музыку и импульсные сигналы.

Информационный параметр по шлейфу сегмента

Информационный параметр по шлейфу сегмента позволяет делить аудиосигналы по признаку речь/музыка/шум [5]. Для расчета данного информационного параметра сегмент длительностью 2 с делится на кадры длиной 32 мс, при частоте дискретизации равной 16 кГц длина каждого кадра составляет N = 512 отсчетов. Для каждого кадра вычисляется спектр $s_i(k)$ с помощью быстрого преобразования Фурье. Затем для каждого кадра рассчитывается параметр потока по следующей формуле:

$$Mflux_i = \sqrt{\frac{dif_i}{sum_i}},$$

где $dif_i = \sum_{k=L}^{H} (s_i(k) - s_{i+1}(k))^2$, $sum_i = \sum_{k=L}^{H} (s_i(k) + s_{i+1}(k))^2$, $L = 16$, $H = 256$.

Для модифицированных значений параметра потока строится гистограмма, из которой высчитывается информационный параметр шлейфа по сегменту с помощью формулы [5]: $T(M) = \sum_{i=M}^{i\_max} H_i$, где H_i – значение гистограммы для *i*-го столбца; M – номер столбца, соответствующий началу хвоста гистограммы; *i_max* – значение общего количества столбцов. Экспериментально установлены следующие значения [5]: $M = 10$, *i_max* = 20.

Для принятия решения речь/музыка/шум устанавливаются диапазоны:

$T < Tmusic$,

$Tmusic < T < Tspeech$,

$Tspeech < T$.

В диапазоне неопределенности между музыкой и речью нужно рассматривать другие информационные параметры, которые уточняют одно из значений речь/музыка/шум. Установлено, что уровень разделения различия речи от музыки может быть выбран равным *Tspeech* = 0,09. Важной особенностью параметра шлейфа является стабильность. Так, дополнение шума к речи сигнализируется уменьшением значения параметра шлейфа, но уменьшение является довольно медленным.

Заключение

Каждый отдельный информационный параметр не позволяет с достаточной уверенностью проводить классификацию сигналов на речь и не речь. Для обнаружения речи в сигналах предлагается использовать все информационные параметры, в той последовательности, как они представлены в работе. Проверка алгоритмов детектирования речи для каждого информационного параметра проводилась на тестовых сигналах, которые были разного вида и содержали только один вид сигналов. Тестовые сигналы были следующих видов «белый шум», хлопки, стуки, шум транспорта, шелест листвы, музыка. Данная проверка позволила уточнить пороговые уровни и выработать требования к созданию алгоритма детектирования речи в реальном режиме времени.

DETECTION OF SPEECH IN REAL-TIME SIGNAL

D.A. BARYSEVICH, H.V. DAVYDAU, V.A. PAPOU

Abstract

Information parameters of the speech signals, allowing to allocate sites of processed information with speech signals in real time, is considered. Modeling in MatLab is executed to detect speech signals in real time.

Список литературы

1. *Paul Alexander Barrett*. Voice activity detector / Патент США № 6061647.
2. *Prabhat K. Gupta, Shrirang Jangi, Allan B. Lamkin et. al.* Voice activity detector for speech signals in variable background noise / Патент США № 5649055.
3. *Чистович Л.А., Венцов А.В., Гранстрем М.П. и др.* Физиология речи. Восприятие речи человеком. Ленинград, 1976.
4. *Douglas J. Nelson, David C. Smith, Jeffrey L. Townsend*. Voice activity detector / Патент США № 6556967.
5. *Sergei N. Gramnitskiy*. Method and system for distinguishing speech from music in digital audio signal in real time / Патент США № 7191128.