УДК 004.65

Д. С. Канищев

ПРИМЕНЕНИЕ ВЕЙВЛЕТ-МЕТОДА ДЛЯ СЕГМЕНТАЦИИ ФОНЕМ ПРИ РАСПОЗНАВАНИИ РЕЧИ

Сегментация фонем при распознавании речи является важной составляющей в процессе распознавания речи. Целью данной статьи является рассмотрение применения одного из подходов к распознаванию фонем – подхода, основанного на вейвлет-преобразовании. Вейвлет-преобразование – преобразование, которое представляет собой свертку вейвлет-функции с сигналом, оно переводит сигнал из временного представления в частотно-временное. Результатом является алгоритм сегментации фонем.

Ключевые слова: распознавание речи, сегментация фонем, вейвлет-преобразование.

Информационные технологии влияют на повседневную жизнь человека все сильнее и сильнее, в связи с чем особенно остро встает проблема взаимодействия между человеком и обрабатывающими информацию устройствами. Хотя в основном под этим взаимодействием понимается использование клавиатур и экранов, не стоит забывать о наиболее распространенном, привычном и быстром способе взаимодействия для человека – речи. Несмотря на последние достижения в области распознавания речи, в основном применяемые в мобильных устройствах, можно сказать что возможности устройств по распознаванию все еще далеки от привычного нам варианта разговора между двумя людьми.

В большинстве подходов к проблеме распознавания речи обрабатываемая голосовая запись должна быть разбита на сегменты перед самим распознаванием. Свойства звукового сигнала в каждом сегменте используется как характеристика отдельного речевого сегмента.

Наиболее часто применяемый метод для этого – это использование разбиения сигнала на временные промежутки, например, по 25 миллисекунд. Достоинства данного подхода – простота реализации и легкость в сравнении сегментов одинаковой длины, с другой стороны – фонемы бывают различной длины, и это нельзя игнорировать. Применение такой сегментации чревато потерями информации из-за соединения разных звуков в один сегмент из-за невозможности использовать длину фонем.

Исходя из недостатков предыдущего подхода можно сказать, что желательно было бы найти такой способ сегментации, который бы определял границы фонем на основе параметров звукового сигнала. Предлагались различные методы [1], но они в основном опираются на звучание конкретных фонем. Такие методы должно быть подстроены под каждый набор обрабатываемых данных и не могут применяться без самого распознавания. Также возможно применение нейронных сетей [2], однако этот подход требует времени на обучение.

Использование спектрального анализа очень эффективно в плане извлечения информации из звукового сигнала. Дискретное вейвлет-преобразование (ДВП) успешно используется во многих приложениях для распознавания речи [3; 4; 5] для спектрального анализа сигналов. В этих приложениях оно применяется для увеличения точности извлечения параметров. Анализ мощности в различных частотных поддиапазонах дает прекрасную возможность выделять начало и конец фонем. На большинстве границ нет заметного падения мощности, с другой стороны сами фонемы на своем протяжении демонстрируют быстрые изменения некоторых поддиапазонов, благодаря которым можно выделить их начальные и конечные точки.

Слуховой аппарат человека в одном из первых шагов анализа звука опирается именно на частотные параметры [1], это было одной из причин данного исследования – вейвлет-преобразование принадлежит к группе частотных преобразований, соответственно, существует возможность найти такие параметры звука голоса, которые важны для человека.

Вейвлет-преобразование можно представить в форме дерева. Корнем дерева будут значения, полученные в результате обработки изначального сигнала. Следующий уровень дерева – еще один шаг ДВП. Последующие уровни строятся рекурсивным применением вейвлет-преобразования для того, чтобы разбить сигнал на общие и детальные части. Рассмотрев различные варианты, было принято остановиться на шести уровнях, которые будут покрывать всю полосу частот человеческого голоса. Список используемых уровней: 86 Гц – 172 Гц, 172 Гц – 345 Гц, 345 Гц – 689 Гц, 689 Гц – 1378 Гц, 1378 Гц – 2756 Гц, 2756 Гц – 5512 Гц. Использование различных вейвлетов показывает совсем небольшие различия в эффективности, однако рекомендуется использовать вейвлет Мейера или симлеты из-за их симметричности.

Изначально ожидалось, что абсолютные значения скорости изменения мощности будут большими в начале и конце фонемы. Но большие абсолютные значения означают не только точки начала и конца. Во-первых, мощность может нарастать некоторое время в начале фонемы. Во-вторых, возможны быстрые изменения мощности в середине сегмента. Лучшим методом определения границ будет использование переходов мощности между поддиапазонами.

В результате экспериментов был разработан следующий алгоритм выделения границ сегментов:

1) Нормализация звукового сигнала голоса путем деления на максимальное значение
2) Раскладывание сигнала на шесть уровней ДВП
3) Вычисление мощностей во всех частотных поддиапазонах
4) Вычисление огибающих в каждом поддиапазоне путем определения наибольших значений
5) Вычисление предельного значения, в рамках которого будут укладываться возможные границы фонемы
6) Определение таких индексов, расстояние между которыми не больше некоторого значения

7) Усреднение значения таких индексов с каждого поддиапазона

В ходе испытаний выявленного алгоритма производилось сравнение автоматической сегментации и ручной 50 слов. Ручная сегментация – это процесс, не отличающийся особой точностью из-за несовершенства человеческого слуха. К тому же фонемы часто перекрываются соседними фонемами. Причина этого заключается в том, что звонкие производятся за счет модуляции воздушного потока из легких путем вибрации голосовых связок, но эта модуляция реагирует на изменения в колебаниях с задержкой. Из-за этого в некоторых образцах возникает некоторая неопределенность именно там, где фонема начинается и заканчивается.

Качество сегментации может оцениваться по двум критериям. Во-первых, число выделенных сегментов должно соответствовать числу фонем слова. Ошибка может быть вычислена следующим образом:

$$\varepsilon_n(w) = \frac{|n_a - n_h|}{n_h}, \text{ где} \quad (1)$$

n_a – количество сегментов, определенных автоматическим образом, а n_h –количество фрагментов, определенных вручную. По результатам экспериментов ошибка составила 23% для слова.

Во-вторых, точность в определении положения границ начала и конца фонемы. Она определяется на основании близости автоматически полученных границ к границам, выделенным вручную (2).

$$\varepsilon_p(w) = \sum_i |p_i - q_i|, \text{ где} \quad (2)$$

p_i – позиция i-границы, определенной автоматически, а q_i – позиция i-границы определенной автоматически. По результатам экспериментов ошибка составила 3.5 миллисекунд на слово.

Используя выделение границ простым разбиением на временные промежутки, мы теряем часть информации, необходимой для качественного распознавания речи. Эффективный и быстрый алгоритм сегментации фонем позволил бы повысить эффективность речевого распознавания в целом, однако проведенное исследование показывает, что использование данного подхода влечет за собой

достаточно высокий уровень ошибок и необходимы дальнейшие улучшения данного алгоритма.

Список литературы

1. *Wang D., Narayanan S.* Piecewise linear stylization of pitch via wavelet analysis. Proc. of Interspeech, 2005.

2. Grayden D. B., Scordilis M. S. Phonemic segmentation of fluent speech. Proc. of ICASSP5, 1994.

3. *Deviren M., Daoudi K.* Frequency and wavelet filtering for robust speech recognition. Joint International Conference on Artificial Neural Networks (ICANN)/International on Neural Information Processing (ICONIP). Istanbul, 2002.

4. *Farooq O., Datta S.* Wavelet based robust subband features for phoneme recognition. IEE Proceedings: Vision, Image and Signal Processing, 151(3):187–193, 2004.

5.. Gowdy J. N., Tufekci Z. Mel-scaled discrete wavelet coefficients for speech recognition. Proc. of ICASSP. Istanbul, 2000.

6. Suh Y., Lee Y. Phoneme segmentation of continuous speech using multi-layer perceptron. In ICSLP 96, 2006.

7. Taft, Marcus and Kenneth I. Forster. "Lexical Storage and Retrieval of Polymorphemic and Polysyllabic Words." Journal of Verbal Learning and Verbal Behavior, 2014.

КАНИЩЕВ Даниил Сергеевич – аспирант кафедры автоматики и телемеханики, Вятский государственный университет. 610000, г. Киров, ул. Московская, 36.

KANISCHEV Daniil Sergeevich – postgraduate student of Department of automation and telemechanics, Vyatka State University. 36 Moskovskaya str., 610000, Kirov.

E-mail: adelantekang@gmail.com