BULLETIN OF NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
ISSN 1991-3494
Volume 2, Number 354 (2015), 286 – 290

ONE OF THE METHODS OF SEGMENTATION OF SPEECH SIGNAL ON SYLLABLES

O. Zh. Mamyrbayev[1], M. M. Kunanbayeva[2], K. S. Sadybekov[2], A. U. Kalyzhanova[2], A. Zh. Mamyrbayeva[3]

[1] The institute of information and calculable technologies MES RK, Almaty, Kazakhstan;
[2] Kazakh National Technical University named after K. I. Satpayev, Almaty, Kazakhstan;
[3]School N 182, Almaty, Kazakhstan.
E-mail: mado_89.89@mail.ru, kuan_91.91@mail.ru

Key words: speech recognition, speech segmentation, zero cross rate.

Abstract. At speech recognition one of the problems of speech segmentation is solved. For the segmentation of the speech signal boundaries between syllables are searched. As an example of the speech signal Kazakh speech is taken. The main features and characteristics of the Kazakh language for recognition are considered. This article describes the algorithm and the method of segmentation of the speech signal based on the syllabic peak, where the energy of the signal reaches the largest value for the boundary syllables.

УДК 004.4

ОДИН ИЗ МЕТОДОВ СЕГМЕНТАЦИИ РЕЧЕВОГО СИГНАЛА НА СЛОГАХ

О. Ж. Мамырбаев[1], М. М. Кунанбаева[2], Қ. С. Садибеков[2], А. У. Калижанова[2], А. Ж. Мамырбаева[3]

[1] Институт информационных и вычислительных технологий МОН РК, Алматы, Казахстан;
[2] Казахский национальный технический университет им. К. И. Сатпаева, Алматы, Казахстан;
[3] Школа № 182, Алматы, Казахстан

Ключевые слова: распознавания речи, сегментация речи, переход уровня сигнала через ноль.

Аннотация. При распознавании речи решается одна из задач сегментация речи. Для сегментации речевого сигнала выполняется поиск границы между слогами. На примере речевого сигнала применяется казахская речь. Рассматривается основные параметры и характеристики казахского языка для распознавания. В данной статье рассмотрены алгоритм и метод сегментации речевого сигнала на основе слогового пика, где энергия сигнала достигает самого большого значения для получения границы слогами.

Введение. Сегментация речевого сигнала является одной из важнейших задач в области информатики и информационных систем для обработки и распознавания речи. Сегментация речевого сигнала необходима для выделения характерных признаков голоса диктора на определённых сегментах речевого сигнала и восстановления формы речевого тракта по акустическому признаку, которая может быть использована синтезе речи по входному тексту и распознавании речи.

В исследованиях можно использовать ручную сегментацию речи, но ручная сегментация речи замедляет работу и практически невозможно точно воспроизвести результаты ручной сегментации, допускает много ошибок при распознавании речи [1].

В информационных системах распознавания речи для сегментации речевого сигнала важным является:

– выделение основных элементов (слов, слогов, фонем) распознавания речи;

– точность сегментации имеет большое влияние наоптимального распознавания речи.

Существует несколько основных типов автоматической сегментации речевого сигнала. К одному из типов относится сегментация речи при условии, что известна последовательность фонемы данной фразы, но результаты распознавания часто ненадежны, а наличие транскрипции возможно только на этапе обучения лексических моделей.

Другой тип не использует априорной информации речи, при этом границы сегментов речи определяются по степени изменения акустических характеристик речевого сигнала. При автоматической сегментации желательно использовать только общие характеристики речевого сигнала, поскольку обычно на этом этапе нет конкретной информации о содержании речи [2].

Для простого сегментации речевого сигнала на паузы и речи, существует метод «blind» segmentation. Данный метод основан на величине и скорости изменения определенных акустических характеристик – это коэффициент перехода уровня сигнала через ноль (ZeroCrossRate) и мера спектрального перехода (SpectralTransitionMeasure), но эксперименты показывают, что для надежной сегментации этих величин недостаточно.

Основы казахской речи образования. Казахский язык входит в кыпчакскую подгруппу тюркских языков (татарский, башкирский, карачаево-балкарский, кумыкский, караимский, ногайский). Вместе с ногайским, каракалпакским и карагашским языками относится к кыпчакско-ногайской ветви [2].

Слова в казахском языке образуются посредством последовательного присоединения к корню или основе слов-аффиксов; грамматических суффиксов и окончаний [2].

Алфавит казахского языка основан на кириллице и состоит из 42 букв, 10 из которых – ә, і, ы, е, ү, ұ, ғ, қ, ң, h – являются специфическими. Буквы в, ф, ц, ч, ь, ъ, е, э используются только при написании слов иноязычного происхождения.

Как и в любом языке, фонетический строй казахского языка включает в себя гласные и согласные звуки. Согласные, в свою очередь, делятся на сонорные, звонкие и глухие. В связи с этим существуют законы сингармонизма, ассимиляции. Суть закона сингармонизма в следующем: в зависимости от слогообразующего гласного в корне слово может принять только твердые или только мягкие аффиксы: /әже-лер/, /бала-лық/, /оқу-шы-лар-ға/.

Этому закону не подчиняются аффиксы принадлежности -/дікі, /-тікі, -/нікі: ата-нікі, қыз-дікі (девушки), Мұраттікі (Мурата) и окончание инструментального падежа:/-мен /-бен, /-пен: Марат-пен, автобус-пен, қыз-бен (с девушкой).

Явление ассимиляции делятся на 2-х типов: прогрессивная и регрессивная. По прогрессивной ассимиляции последующий согласный звук на слоговой границе употребляется предыдущему. Например: кітап-тар, қалам-дар, т.е. к словам, оканчивающимся на глухие и звонкие б, в, г, д прибавляются аффиксы, начинающиеся на глухие согласные: а слова со звонкими, сонорными и гласными в конце требуют аффиксов со звонкими или сонорными звуками: аға-дан, қыз-дың, үй-дің.

Регрессивная ассимиляция предполагает озвучивание глухих согласных «қ, к п» в конце слова если прибавляемые аффиксы начинаются на гласные. Например: кітап-кітабым (моя книга), оқулық, (учебник), оқулығы (его учебник) [2].

Классификация звуков

Виды	Звуки	Специфические звуки
1. Гласные звуки	Сложно- сочиненные: а, ө, о, е. Монофтонг: ұ, ү, ы, і. Дифтонг: и, у. Введенные с русского языка: э, ё, ю, я	ә, ө, і, ұ, ү
2. Согласные звуки	б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, ф, х, ц, ч, ш, щ, h, (у), знаки: ь, ъ	қ, ғ, ң, h

В отличие от русского языка, существительные в казахском языке не имеют категорию рода, поэтому нет согласования между существительным и прилагательным, существительным и числительным. При склонении сочетаний из этих частей речи падежные окончания прибавляются к существительным.

Слог – это гласный звук или несколько звуков в слове, которые в процессе произношения произносятся одним толчком воздуха. Слоги, состоящие из двух и более звуков, могут оканчиваться либо на гласный:

1) открытые слоги: ана – мать либо на согласные;

2) полузакрытые слоги: от –огонь, өрт –пожар, либо начало и конец слога закрыты согласными;

3) закрытые слоги: тас – камень, кен – руда.

Ударение в казахском языке по сравнению с русским более постоянное падающее на один определенный слог слова, обычно последний. Если к словам прибавляются аффиксы, то и ударения в них передвигаются на последние слоги аффиксов

Математическая подстановка сегментации речевого сигнала на слоги. Входящий речевой сигнал записывается в виде последовательности отчетов y_i …

$$Y = y_0, y_1, \dots, y_i, \dots; \text{ где } i = 0,1,2, \dots$$

Последовательность речевого сигнала разделяется на фреймы длиной 128 отсчета (соответственно (128·1000)/11025 ≈ 11 мс). Размер фрейма позволяет точно определит границы между слогами [3].

По следующей формуле находим среднее значение энергии во фрейме речевого сигнала длиной 128 отсчета:

$$E_i = \frac{\sum_{j=i*128}^{i*128+127} y_j^2}{128}; \text{ где } i = 0,1,2, \dots \quad (1.1)$$

Полученные значения по формуле (1.1) являются средней энергией короткого времени на промежутке 11 мс. Подсчитаем среднее значение энергии короткого времени трех соседних участков по формуле:

$$E_i^* = \frac{E_i + E_{i+1}}{2}; \text{ где } i = 0,1,2, \dots \quad (1.2)$$

Таким образом, рассчитываем среднюю энергию для фреймов 2·128 = 256 отсчета. Фреймы берутся с наложением и сдвигом соседних интервалов на 128 отсчета (рисунок 1).

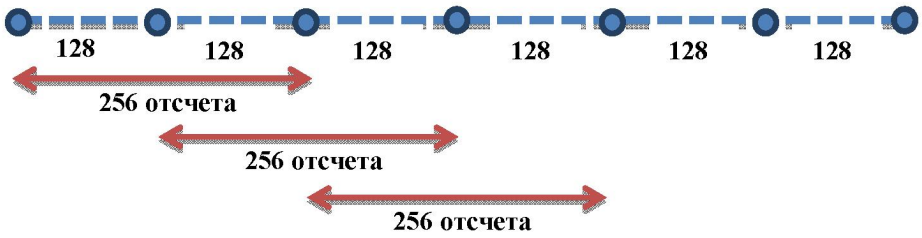

Рисунок 1 – Разделение речевого сигнала на фреймы

Основной тон казахского языка меньше, чем 256 / 11025 = 0.023 сек., что соответствует основной частоте I / 0.023 = 75,5 Гц. Поэтому, энергия фрейма длиной в 256 отсчета заключает энергию, по крайней мере, одного периода основного тона. Таким образом, из последовательности речевого сигнала $Y = y_0, y_1, \dots, y_i, \dots$; где $i = 0,1,2, \dots$ рассчитаем последовательности средней энергии участков в 192 отсчетов $E^* = E_1^*, E_2^*, \dots, E_i^*, \dots$.

Каждый слог имеет слоговой пик, где энергия сигнала достигает самого большого значения.

Между двумя слоговыми пиками имеется точка, соответствующая граница, которая разделяет слоги [3].

Алгоритм определения границы между слогами. Для определения точки-границы между двумя слогами применяется следующий алгоритм:

1. Определения слоговых пиков.
2. Определение точки наименьшей энергией между слоговыми пиками.

В большинстве случаев эта точка является границей между двух слов. Но есть случаи, когда эта точка была конечной точкой шипящего согласного следующего слова. Так что надо определить слева от этой точки был ли согласный шипящий или голосовой. Определим шипящих реализованных по числу переходов речевого сигнала через нуль. Подсчитаем долю числа переходов через нуль для участка длиною в N отсчетов, который находится слева от точки с минимумом энергии и заканчивается отсчетом m:

$$Z_x(m) = \frac{1}{N}\sum_{n=m-N+1}^{m}\frac{|sgn(x_n - sgn(x_{n-1})|}{2}, где\ N = 256;$$

$$sgn(x) = \begin{cases} 1\ если\ x > 0 \\ 0\ если\ x = 0 \\ -1\ если x < 0 \end{cases}$$

При принятой частоте дискретизации 11 025 отсчетов в секунду число переходов сигнала через нуль у щелевого звука всегда больше 14 на 100 отсчетов. Соответственно переходов через нуль ≥ 0,14, а у голосового сигнала меньше этого числа.

Интервал наблюдения равен 100*Δt = $\frac{100}{11025}$ [*сек*] , число переходов через нуль 14, так как в среднем на 1 период приходится 2 перехода, то в интервале $\frac{100}{11025}$ [*сек*] укладывается $\frac{14}{2}$ периодов.

Если пропорция $\frac{Z_x(m)*100}{256}$ больше 14, то считаем участок щелевым и считаем переходов на предыдущем участке. Выполнение этого условия определяет отправную начальную точку шипящего согласного.

3. Проверка числа переходов через нуль слева от точки с минимальной энергией, чтобы точно определить точку границы между слогами.

И это будет точное разделения двух слогов, соответствующих двум словам [3].

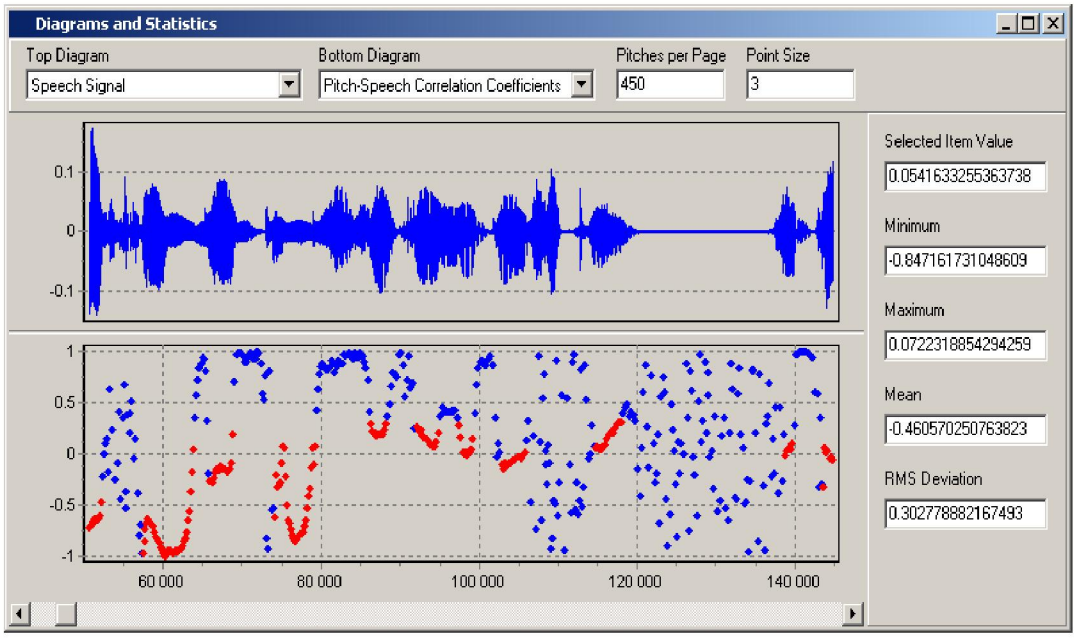

Рисунок 2 – Результат автоматической сегментации речевого сигнала по слогам

В статье рассматривается, что предложенный метод успешно разбивает речевой сигнал на непересекающиеся вокализованные сегменты и может быть применен к различным задачам.

ЛИТЕРАТУРА

[1] Сорокин В.Н. Сегментация речи на кардинальные элементы // Информационные процессы. – 2006. – С. 177-207.

[2] Жиляков Е.Г. Методы обработка речевых данных в информационно-телекоммуникационных системах на основе частотных представлений. – М., 2000. – С. 456-460.

[3] Фирсова А.А. О различиях распределения энергия звуков русской речи и шума. – М., 2010. – С. 204-207.

REFERENCES

[1] Sorokyn V. *Segmentation of speech on cardinal elements.* Informative processes **2006,** 177-207. (in Russ)

[2] Zhylyakov E. Methods of processing of speech data in the information-telecommunication systems on the basis of frequency presentations. M., **2000.** 456-460. (in Russ)

[3] Firsova A. About distinctions of distribution of sounds' energy of Russian speech and noise. Moscow, **2010,** 204-207. (in Russ)

СИГМЕНТТІ СӨЗДЕРДЕ БУЫНДЫ БӨЛУДІҢ БІР ӘДІСІ

О. Ж. Мамырбаев, М. М. Кунанбаева, Қ. С. Сәдібеков,
А. У. Калижанова, А. Ж. Мамырбаева

Тірек сөздер: сөздік сигнал, сөзді сегменттеу, сигналдың нөлдік деңгейінен өтуі.

Аннотация. Сөзді тану кезеңінде сөзді сегменттеу есебі шешіледі. Сөздік сигналды сегменттеу кезінде буындардың арасындағы шекараны іздеу іске асырылады. Мысал ретінде қазақ тіліндегі сөздік сигнал алынған. Тану үшін қазақ тілінің негізгі параметрлері мен мінездемелері қарастырылады. Мақалада сигнал энергиясының ең үлкен мәнін, буындар арасындағы шекараны анықтау үшін, буындар арасындағы шың негізінде сөздік сигналды сегменттеу әдісі мен алгоритмі қарастырылады.

Поступила 20.03.2015 г.