BULLETIN OF NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
ISSN 1991-3494
Volume 3, Number 367 (2017), 137 – 143

S. S. Nesipova, L. B. Ilipbayeva

Kazakh National Research Technical University named after K. I. Satpayev, Almaty, Kazakhstan.
E-mail: nesipova_s@mail.ru, ilizat1011@mail.ru

NEURAL NETWORK APPLICATION IN SPEECH PROCESSING

Abstract. The article describes the methods of preliminary processing of the speech signal. The process of obtaining a 24-element spectral vector, as spectral analysis is one of the commonly used parametric representations of speech. It was simulated the model of pre-processing the speech signal in Matlab. As a result of the preliminary treatment the illustration of a single block of speech signal, its circuit after processing by the filter of the first order, after application of Hamming window, the amplitude value of the fast Fourier transform and the values of the vector whose components obtained after averaging the amplitude values, were obtained. The calculated values of Mel-frequency cepstral coefficients were used to generate the feature vector. Also a model for recognition of speech signals based on neural network algorithm of Kohonen was proposed.

Key words: speech recognition, frequency of basic tone, Kohonen neural network, preprocessing of the speech signal.

УДК 004.032.26

С. С. Несипова, Л. Б. Илипбаева

НАО Казахский национальный исследовательский технический университет им. К. И. Сатпаева, Алматы, Казахстан

ПРИМЕНЕНИЕ НЕЙРОННОЙ СЕТИ В ОБРАБОТКЕ РЕЧЕВОГО СИГНАЛА

Аннотация. В статье описаны методы предварительной обработки речевого сигнала. Исследован процесс получения 24 элементного спектрального вектора, так как спектральный анализ является одним из часто используемых параметрических представлений речи. Была смоделирована модель предварительной обработки речевого сигнала в среде Matlab. В результате предварительной обработки были получены иллюстрации одного блока речевого сигнала, её схема после обработки его фильтром первого порядка, после применения окна Хэмминга, амплитудные значения быстрого преобразования Фурье и значения вектора, компоненты которого получены после усреднения амплитудных значений. Рассчитанные значения мелчастотных кепстральных коэффициентов были использованы для формирования вектора признаков. Также была предложена модель распознавания речевого сигнала на основе нейронной сети по алгоритму Кохонена.

Ключевые слова: распознавание речи, частота основного тона, нейронная сеть Кохонена, предварительная обработка речевого сигнала.

В современном мире информационно-телекоммуникационных систем одной из актуальных проблем, которая является не до конца решенной остается задача обработки речевых данных. К наиболее распространенным среди них относятся системы идентификации по голосу, преобразование речи в текст, синтез по тексту и голосовое управление. При исследовании особенностей распределения энергии звуков было выявлено, что все звуки имеют индивидуальное распределение энергии по частотным интервалам. К тому же, распределение энергии зависит от местоположения распространителя звука, диктора, его эмоционального состояния и интонации. Среди характе-

ристик речевых сигналов можно выделить те, которые незначительно изменяются на протяжении всего звука. Одним из таких параметров, который характеризует частоту колебаний голосовых связок, является частота основного тона [1]. Частота основного тона меняется во время разговора человека и её относительное изменение может достигать 15%. В европейских языках основной тон передает эмоциональную составляющую речи, а в некоторых восточных – смысловую. Экспериментально установлено, что для женских голосов период основного тона составляет в среднем от 220 до 350 Гц, а для мужских от 100 до 220 Гц. Колебание связок является одним из основных параметров источника голосового возбуждения речевого тракта. Они придают голосу звучание и характеризуют его высоту [2]. Частота основного тона зависит от длины связок, их массы и натяжения [3]. Для приближенного понимания этой связи можно представить струны: чем длиннее и тяжелее складки (эти свойства - врожденные), тем более низкий тон имеет голос, чем складки короче и тоньше – тем голос выше.

Предварительная обработка речевых сигналов производится для получения множества спектральных векторов, которые характеризуют этот сигнал.

Так как спектральные характеристики речевого сигнала относительно постоянны на интервале в несколько десятков миллисекунд в современных распознавателях он рассматривается как стационарный. Поэтому, основной целью предварительной обработки входного речевого сигнала является разбиение сигнала на интервалы и получение для каждого интервала сглаженной спектральной оценки.

Для предварительной обработки берется типичная величина одного интервала со значением 25,6 мс, соседние же интервалы берутся со смещением относительно предыдущего интервала. Применяемая величина перекрытия интервалов равна 10 мс. Результатом предварительной обработки каждого из указанных интервалов является вектор из нескольких десятков спектральных значений.

Предварительная обработка речевого сигнала состоит из следующих этапов:

1. Оцифрованный, то есть дискретизированный во времени и квантованный по уровню речевой сигнал разбивается на блоки с интервалом 25,6 мс со смещением каждые 10 мс, то есть, блоки располагаются по 409 отсчетов каждый со смещением на 160 отсчетов.

2. Рассеивание губ приводит к ослаблению сигнала и для его компенсирования применяют высокочастотное усиление посредством пропускания сигнала через фильтр первого порядка

$$S(1) = 0; S(n) = y(n) - y(n-1), n = 2,...,409,$$

где $y(n)$ – n--й отсчет в блоке.

3. Возможность свертки анализируемого участка сигнала с оконной функцией Хэмминга [4] реализована для устранения явления просачивания спектральных составляющих в оболочке, которая определяется согласно выражению

$$D(n) = (0{,}54 - 0{,}46 \cdot \cos(2\pi \cdot (n-1)/408)) \cdot S(n) \text{ для } n = 1,..., 409.$$

4. Для получения спектральных оценок используется дискретное преобразование Фурье:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi nk/N}, k = \overline{0,\ N-1},$$

где $x_n, n = \overline{0, N-1}$ - дискретный сигнал, N – период преобразования (или количество преобразуемых отсчетов сигнала). За счет дополнения его справа нужным количеством нулей длина блока увеличивается до 512 элементов. После применяется быстрое преобразование Фурье (БПФ) длиной 512 точек, и на выходе получаем 512 спектральных комплексных значений. Так как, 512 значений, к которым применяется преобразование Фурье, являются действительными, то полученные спектральные комплексные значения попарно сопряжены, то есть второе значение с 512-м, третье – с 511-м и т.д. Так как последние 256 комплексные значения комплексно сопряжены с предыдущими и не несут новой информации их преобразование игнорируются.

5. Для начальных 256 комплексных спектральных значений находятся их амплитуды. В пределах «треугольных» частотных полос амплитудный спектр Фурье сглаживается (усредняется)

добавлением амплитуд спектральных коэффициентов, которые располагаются на нелинейной (подобной логарифмической) Mel-шкале. Для предельной частоты языка равной 16 кГц берут 24 таких частотных полосы.

Метод Mel-шкалы (Mel Frequency Cepstral Coefficient - MFCC) основан на модели функционирования органов слуха человека и использует частотную шкалу мел. Эта мел шкала моделирует частотную чувствительность человеческого уха, которая является линейной до 1000 Гц и логарифмической в значениях более 1000 Гц [4].

Первый амплитудный коэффициент – постоянная составляющая спектра игнорируется, а амплитуды остальных 255 спектральных значений усредняются. Процесс усреднения реализуется как 24 треугольные полосопропускные фильтры. Нижняя, средняя и верхняя частоты таких полос приведены в таблице.

24-элементный спектральный вектор

Полоса	Нижняя частота, Гц	Средняя частота, Гц	Верхняя частота, Гц
1	0	74,24	156,4
2	74,24	156,4	247,2
3	156,4	247,2	347,6
4	247,2	347,6	458,7
5	347,6	458,7	581,6
6	458,7	581,6	717,5
7	581,6	717,5	867,9
8	717,5	867,9	1034
9	867,9	1034	1218
10	1034	1218	1422
11	1218	1422	1647
12	1422	1647	1895
13	1647	1895	2171
14	1895	2171	2475
15	2171	2475	2812
16	2475	2812	3184
17	2812	3184	3596
18	3184	3596	4052
19	3596	4052	4556
20	4052	4556	5113
21	4556	5113	5730
22	5113	5730	6412
23	5730	6412	7166
24	6412	7166	8000

Каждый треугольный полосопропускной фильтр находит взвешенное среднее спектральное значение, которое соответствует частотам в пределах между нижней и верхней частотами для данного фильтра. Если же амплитуда соответствует точно средней частоте полосы, то она умножается на коэффициент равный единице. При передвижении частоты от середины к нижней или верхней границе коэффициент уменьшается от единицы до нуля. Полученные произведения амплитуд на коэффициенты добавляются и делятся на число амплитудных значений. В результате находится взвешенное среднее значение для данной полосы частот. Таким образом, 256 амплитудам соответствуют частоты от 0 до 8000 Гц, т.е. шаг передвижения равен 8000/256=31,25 Гц. Это означает, что первой амплитуде соответствует частота 0 Гц, второй – 31,25 Гц, третьей – 61,25 Гц.

Например, для первой полосы частот Mel-шкалы нижняя частота равна 0 Гц, средняя частота 74,24 Гц, верхняя частота 156,4 Гц.

Из рисунка 1 видно, что в первую полосу частот попадают первая (0 Гц), вторая (31,25 Гц), третья (62,5 Гц), четвертая (93,75 Гц), пятая (125 Гц) и шестая (156,25 Гц) амплитуды. Согласно рисунку третьей амплитуде соответствует коэффициент равный $62,5/74,24 \approx 0,84$; а коэффициент для пятой амплитуды равен $(156,4\text{-}125)/(156,4\text{-}74,24) \approx 0,38$.

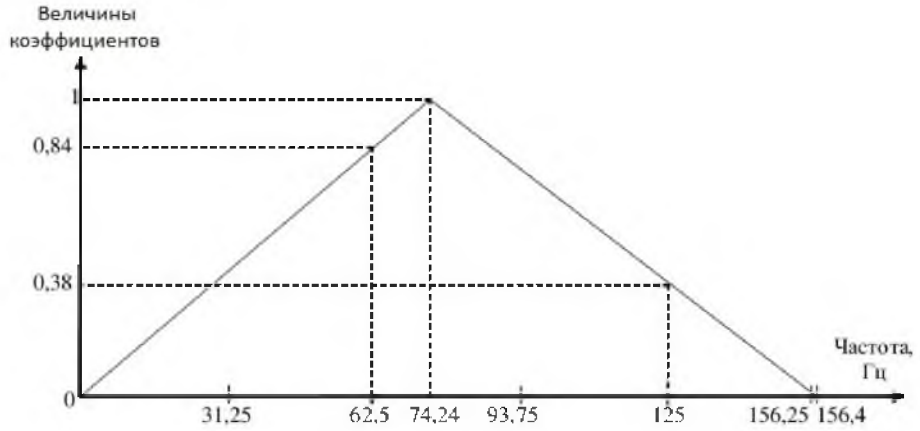

Рисунок 1 – Mel-шкала частот

В результате описанных выше действий получаем 24-элементный спектральный (акустический) вектор.

Для моделирования алгоритма предварительной обработки речевых сигналов выбрана среда Matlab, предоставляющая широкие возможности по обработке речевых сигналов и проведению трудоемких вычислений.

На втором рисунке показан речевой сигнал test.wav, дискретизированный с частотой 16 кГц и разрядностью 16 разрядов.

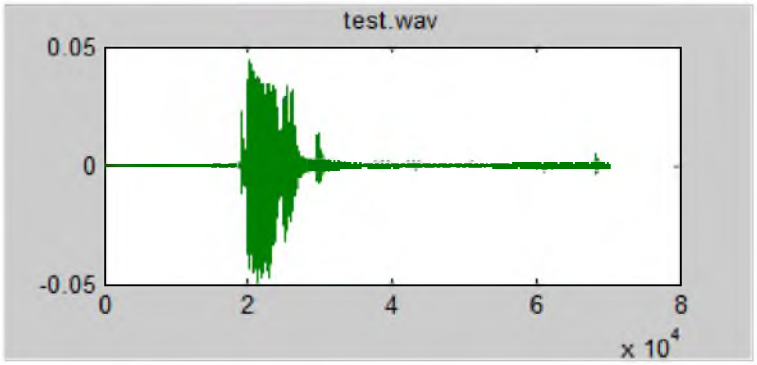

Рисунок 2 – Речевой сигнал test.wav

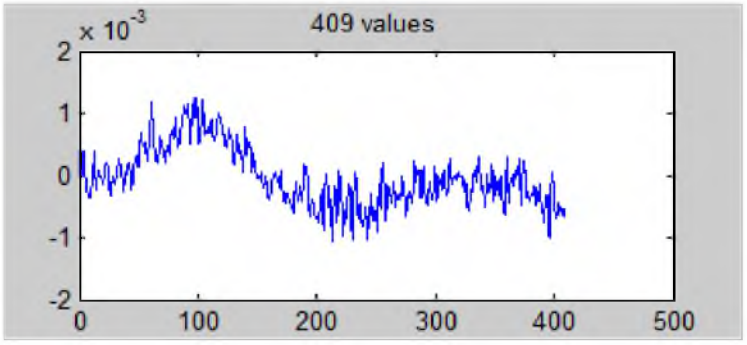

Рисунок 3 – Один блок речевого сигнала

На третьем рисунке показан один блок (интервал) указанного речевого сигнала длительностью 25,6 мс. Этому интервалу соответствует 409 отсчетов.

На четвертом рисунке можно увидеть один блок речевого сигнала после обработки его фильтром первого порядка.

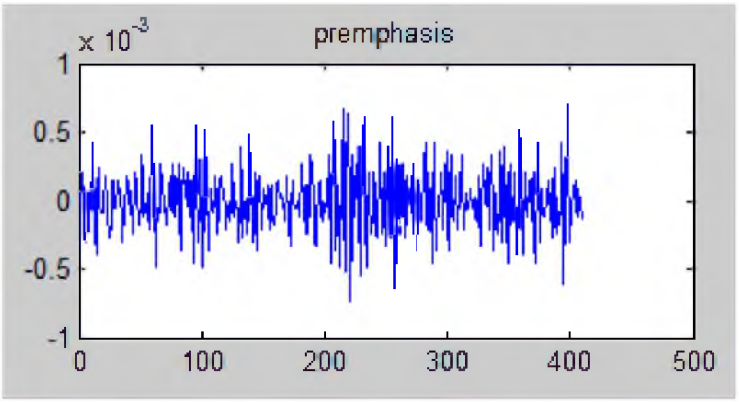

Рисунок 4 – Сигнал, обработанный фильтром первого порядка

Пятый рисунок иллюстрирует один блок после применения окна Хэмминга.

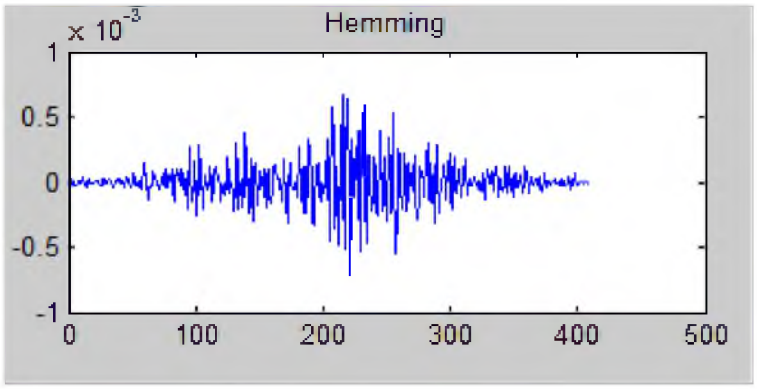

Рисунок 5 – Сигнал, после применения окна Хэмминга

Шестой рисунок дает нам 512 амплитудных значений быстрого преобразования Фурье одного блока речевого сигнала.

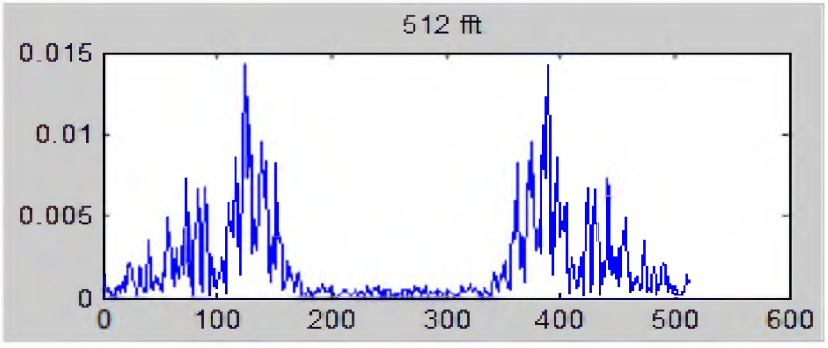

Рисунок 6 – 512 амплитудных значений сигнала

Так как амплитудные значения БПФ совпадают попарно, как было упомянуто выше, то были взяты только первые 256 амплитудных значений, которые показаны на рисунке 7.

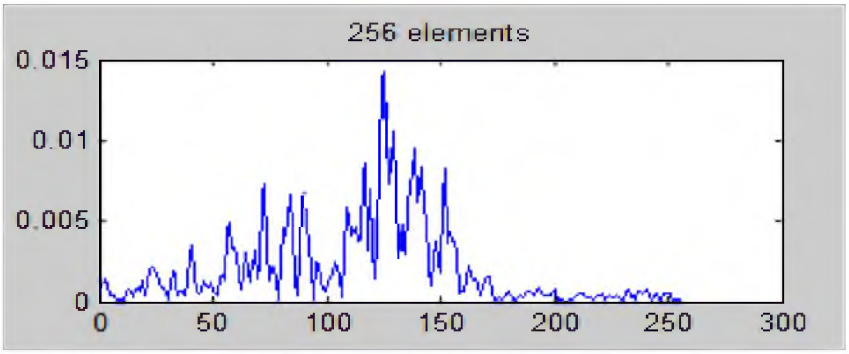

Рисунок 7 – 256 амплитудных значений сигнала

Восьмой рисунок дает значения 24-элементного вектора, компоненты которого получены после усреднения 256 амплитудных значений в пределах 24 «треугольных» частотных полос.

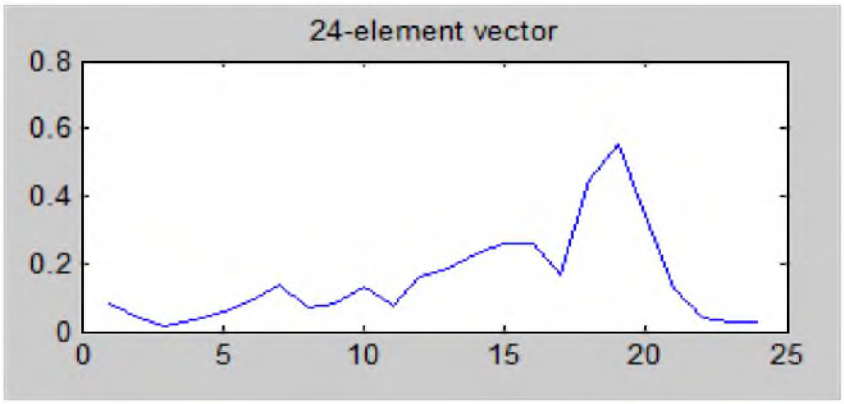

Рисунок 8 – 24-элементный вектор

В результате проделанной работы предложена модель распознавания речевого сигнала на основе нейронной сети Кохонена, пример которой представлен на рисунке 9. Нейронные сети Кохонена помогают реализовать более сложные системы, так как при одновременном использовании в одной системе сетей с различной структуры, можно добиться реализации дикторонезависимой системы.

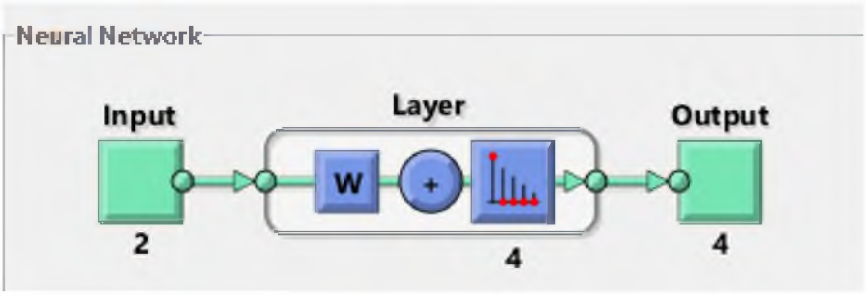

Рисунок 9 – Модель нейронной сети Кохонена

Подобные системы распознавания речевого сигнала, построенные на основе искусственных нейронных сетей, могут найти применение в различных отраслях промышленной и бытовой электроники. Поскольку нейронные сети справляются с задачей распознавания гораздо быстрее и позволяют существенно повысить точность системы распознавания речевого сигнала.

ЛИТЕРАТУРА

[1] Лузин Д.А. Разработка и исследование системы автоматического выделения основного тона речи: Автореферат. – Ижевск, 2009.

[2] Бабкин В.В. Помехоустойчивый выделитель основного тона речи // Труды 7-й Международной конференций и выставки Цифровая обработка сигналов и её применение (DSPA-2005). – М., 2005.

[3] Рабинер Л.Р. Теория и применение цифровой обработки сигналов / Л.Р.Рабинер, Р.В.Шафер. – М.: Радио и связь, 1981. – 496 с.

[4] X.Huang, A.Acero, H.Hon. Spoken language processing: a guide to theory, algorithm, and system development. – Prentice Hall PTR, 2001. – P. 936.

REFERENCES

[1] Luzin D.A. Razrabotka i issledovanie sistemy avtomaticheskogo vydelenija osnovnogo tona rechi: Avtoreferat. – Izhevsk, 2009.

[2] Babkin V.V. Pomehoustojchivyj vydelitel' osnovnogo tona rechi // Trudy 7-j Mezhdunarodnoj konferencij i vystavki Cifrovaja obrabotka signalov i ejo primenenie (DSPA-2005). – M., 2005.

[3] Rabiner L.R. Teorija i primenenie cifrovoj obrabotki signalov / L.R.Rabiner, R.V.Shafer – M.: Radio i svjaz', 1981. – 496 s.

[4] X.Huang, A.Acero, H.Hon. Spoken language processing: a guide to theory, algorithm, and system development. – Prentice Hall PTR, 2001. – P. 936.

С. С. Несипова, Л. Б. Илипбаева

Қ. И. Сатпаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан

АҚПАРАТТЫ СИГНАЛДЫ ӨҢДЕУДЕ НЕЙРОЖЕЛІЛЕРДІ ҚОЛДАНУ

Аннотация. Мақалада сөйлеу сигналының бастапқы өңдеу әдістері жайлы баяндалған. Элемент саны 24 болатын спектралды векторды алу процессі зерттеліп қарастырылған, себебі спектралды анализ сөзді параметрлі ұсынудың жиі қолданылатын әдісі болып табылады. Matlab бағдарламалау ортасында сөйлеу сигналдың бастапқы өңдеу моделдері құрастырылды. Бастапқы өңдеу процесі нәтижесінде сөйлеу сигналының бір блогының суреттемесі, оның бірінші қатарлы сүзгіден өту схемасы, Хэмминг терезесін қолданганнан кейінгі сұлбасы, тез Фурье түрлендіруінің амплитудалық мәндері және амплитудалық мәндерді орташалағаннан кейінгі вектордың мәндері алынды. Есептелген мел-жиілікті кепстралды коэффициенттердің мәндері белгі векторларын қалыптастыруда қолданылды. Сонымен қатар нейрожелісінің Кохонен алгоритімі негізінде сөйлеу сигналды анықтап табудың моделі ұсынылды.

Түйін сөздер: сөздерді ажырату, негізгі екпін жиілігі, Кохонен нейрожелісі, ақпаратты сигналды бастапқы өңдеу.