Теория сигналов и систем

УДК 004.394:534.78

Нейросетевой алгоритм выделения тональных, шумовых и паузных участков устной речи

И.Ю.Бондаренко[1], **О.Н.Ладошко**[2]

[1]Донецкий нацональный технический университет,
ул. Артема, 58, Донецк, 83000, Украина

[2]Национальный технический университет Украины «КПИ»,
пр. Победы, 37, Киев-56, 03056, Украина

Рассматривается проблема автоматического выделения тональных, шумовых и паузных участков устной речи. Для решения этой проблемы предлагается нейросетевой алгоритм, выполняющий классификацию последовательности фреймов, на которые разбивается речевой сигнал. На материале речевых корпусов TIMIT и NTIMIT проведены эксперименты по оценке качества, надежности и скорости работы алгоритма в дикторонезависимом режиме, в том числе в условиях нестационарного шума, вызванного влиянием телефонного канала.

Ключевые слова: *классификация речь-шум-пауза, нейронная сеть.*

Введение

Типичная структура системы автоматического распознавания речи представляет собой последовательность модулей считывания, предварительной обработки и распознавания речевого сигнала. Важную роль в функционировании модуля предварительной обработки речевого сигнала играет алгоритм автоматического выделения тональных, шумовых и паузных участков устной речи. Предварительная классификация участков речевого сигнала на тон (вокализованную речь), шум (невокализованную речь) и паузу (отсутствие речи), во-первых, позволяет более точно вычислить просодические, в частности, тональные, характеристики речи, во-вторых, упрощает процесс дальнейшей классификации речевого сигнала на фонемы и слова.

От алгоритма выделения тональных, шумовых и паузных участков устной речи требуется:

1) функционировать в дикторонезависимом режиме (во многих случаях отсутствует возможность проводить адаптацию системы распознавания речи к голосу конкретного диктора);

2) обеспечивать приемлемую точность выделения участков тона, шума и паузы в сигнале и надёжность в условиях нестационарного шума;

3) работать достаточно быстро, чтобы вся система распознавания речи, использующая данный алгоритм, могла функционировать в реальном масштабе времени.

Существует ряд алгоритмов (см., например, [2] или [5]), использующих сложную систему признаков, таких, как спектральные или мелчастотные кепстральные коэффициенты, и обеспечивающих неплохую точность выделения тоновых, шумовых и паузных участков устной речи. Однако их недостатком является громоздкая процедура вычисления используемой системы признаков, сравнимая по вычислительной сложности с процедурой самого распознавания. В других алгоритмах [1, 8] используются просто вычисляемые системы признаков (энергия сигнала, число переходов через ноль, коэффициенты автокорреляционной функции и т.п.), но применяемые там классификаторы не позволяют обеспечить той точности, которая была бы возможна при использовании более сложных нелинейных классификаторов.

Исходя из вышеописанного, в данной статье была поставлена следующая цель — разработать дикторонезависимый алгоритм выделения тональных, шумовых и паузных участков устной речи, основанный на использовании достаточно простой системы признаков и мощного нелинейного классификатора — многослойной нейронной сети с сигмоидальными функциями активации. Поскольку такая нейронная сеть является универсальным апроксиматором [9], то она позволяет аппроксимировать сколь угодно сложные дискриминативные функции. В то же время нейронная сеть легко отображается на

вычислительные устройства с параллельной архитектурой, что обеспечивает быструю работу любого нейросетевого алгоритма.

1. Описание алгоритма

Речевой сигнал подвергается скользящему оконному анализу с длиной окна 20 мсек и шагом 10 мсек. В результате такого анализа речевой сигнал разбивается на T речевых фреймов. Для каждого фрейма с помощью нейросетевого классификатора определяется принадлежность к одному из классов — тон, шум или пауза.

Входным сигналом нейросетевого классификатора является вектор из трёх компонент — параметров речевого сигнала, вычисляемых на каждом t-м фрейме, $t = 1..T$:

1) E_t — кратковременная энергия сигнала [10];

2) $R1_t$ — отношение 1-го коэффициента автокорреляционной функции сигнала к её 0-му коэффициенту [10];

3) ZCR_t — число переходов через ноль [10].

Пример исходного звукового сигнала, полученного при произнесении диктором-мужчиной английской фразы «At twilight on the twelfth day we'll have Chablis», и трёх вышеперечисленных признаков звукового сигнала, по которым проводится классификация, приведён на рис.1.

Выходным сигналом нейросетевого классификатора является вектор из трёх компонент, определяющих степень принадлежности t-го фрейма одному из трёх классов (тону, шуму или паузе). Если t-й фрейм является тональным фреймом, то выходной сигнал нейронной сети должен принимать значение (+1; –1; –1). Если t-й фрейм является шумовым фреймом, то выходной сигнал нейронной сети должен принимать значение (–1; +1; –1). И, наконец, если t-й фрейм является паузным фреймом, то выходной сигнал нейронной сети должен принимать значение (–1; –1; +1).

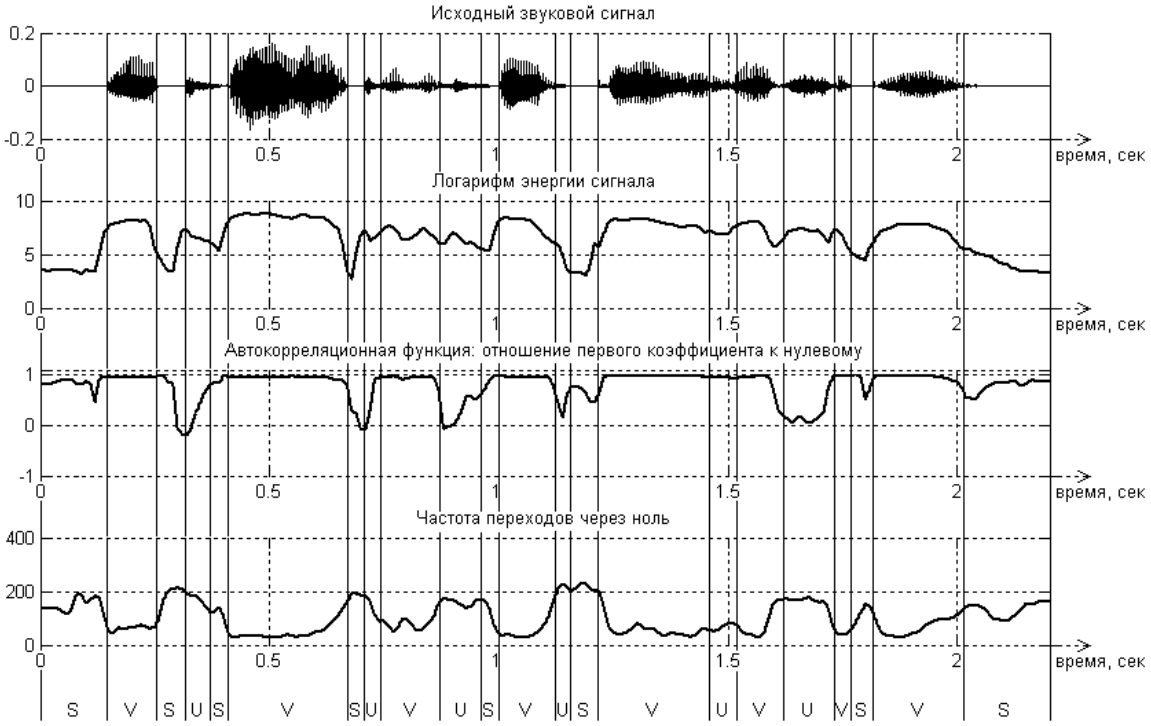

Рис.1. Звуковой сигнал, вручную размеченный на тональные (V – voiced), шумовые (U –unvoiced) и паузные (S – silence) участки, и признаки этого сигнала, используемые для классификации

Нейронная сеть, решающая задачу классификации «тон/шум/пауза», является классической многослойной сетью с полными последовательными связями (см. рис.2). В данной работе используется нейронная сеть с одним скрытым слоем, число нейронов в котором подбиралось экспериментально. В качестве функции активации нейронов используется рациональная сигмоида следующего вида:

$$f(x) = \frac{2 \cdot x}{1 + |x|}.$$

Такая функция активации, во-первых, обеспечивает биполярность всех сигналов внутри сети и тем самым повышает эффективность обучения этой сети [7], а во-вторых, быстро вычисляется (например, в отличие от другой би-

полярной сигмоиды — гиперболического тангенса).

Нейронная сеть обучается с учителем. Обучающее множество формируется на основе списка речевых сигналов и их временных разметок на тональные, шумовые и паузные участки, выполненных вручную. Входные сигналы всех обучающих примеров нормализуются так, чтобы мат.ожидание по всем компонентам входного сигнала было нулевым, а среднеквадратичное отклонение — единичным. Фрагмент подготовленного таким образом обучающего множества, включающий в себя по два обучающих примера для каждого из классов, приведён в табл. 1.

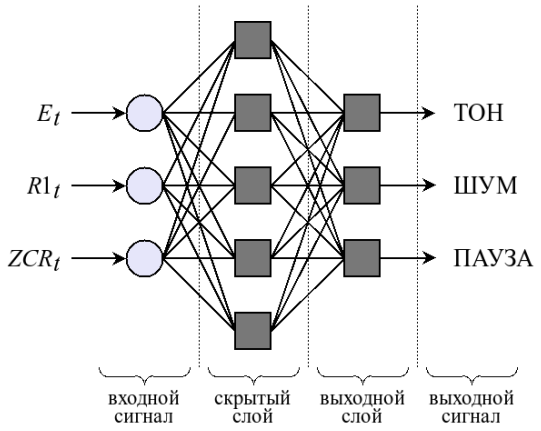

Рис.2. Структура нейросетевого классификатора тональных, шумовых и паузных участков устной речи с пятью нейронами в скрытом слое

В качестве алгоритма обучения используется алгоритм обратного распространения ошибки, функционирующий в режиме «онлайн» [4]. Этот алгоритм является более точным, чем варианты пакетного обратного распространения [6], и более быстрым, чем алгоритмы глобальной оптимизации, такие как алгоритм имитации отжига или генетические алгоритмы [11].

2. Результаты экспериментов

Для определения точности и надёжности работы предложенного нейросетевого алгоритма выделения тональных, шумовых и паузных участков устной речи были проведены эксперименты на материале речевых корпусов TIMIT и NTIMIT.

TIMIT — это классический речевой корпус, содержащий свыше 5 часов цифровых звукозаписей различных английских фраз, произнесённых 630 дикторами на 8 диалектах американского английского. Все звукозаписи имеют временную пофонемную разметку, выполненную профессиональными фонетистами. Речевой корпус разбит на два непересекающихся множества: обучающее и тестовое [7].

Речевой корпус NTIMIT построен на основе речевого корпуса TIMIT. Звукозаписи речевого корпуса TIMIT были пропущены через телефонные каналы американской телефонной компании NYNEX и заново оцифрованы. Это позволило представить в речевом корпусе NTIMIT звукозаписи с нестационарным случайным шумом, характерным для естественного телефонного канала связи [3].

Экспериментальные исследования были проведены следующим образом. Вначале нейронная сеть, выполняющая выделения тоновых, шумовых и паузных участков устной речи, была обучена на материале обучающего множества речевого корпуса TIMIT. Затем было поставлено два эксперимента:

1) нейронная сеть, обученная на обучающем множестве корпуса TIMIT, тестировалась на тестовом множестве этого же корпуса;

2) нейронная сеть, обученная на обучающем множестве корпуса TIMIT, тестировалась на тестовом множестве корпуса NTIMIT.

Целью этих экспериментов было определить, как точно разработанный нейросетевой алгоритм выделяет тоновые, шумовые и паузные участки устной речи, и насколько этот алгоритм надёжен, т. е. насколько снижается точность работы алгоритма в условиях нестационарного шума.

Ошибка классификации «тон/шум/пауза» считалась по следующей формуле:

$$Err = \frac{T_{err}}{T_{all}} \cdot 100\%,$$

где T_{err} — количество неправильно классифицированных звуковых фреймов, а T_{all} — общее количество звуковых фреймов.

Таблица 1. Фрагмент обучающего множества, используемого для обучения нейронной сети решению задачи классификации «тон/шум/пауза»

Входной сигнал			Желаемый выходной сигнал		
E_t	$R1_t$	ZCR_t	Тон	Шум	Пауза
1,434	0,609	–1,071	+1	–1	–1
1,274	0,618	–1,158	+1	–1	–1
–0,232	–1,185	1,076	–1	+1	–1
–0,144	–0,981	1,041	–1	+1	–1
–1,939	0,360	1,024	–1	–1	+1
–2,144	–0,240	0,937	–1	–1	+1

Таблица 2. Распределение ошибок классификации «тон/шум/пауза» по отдельным классам при обучении и тестировании классификатора на материале речевого корпуса TIMIT.

	Ошибка выделения тона	Ошибка выделения шума	Ошибка выделения паузы
Тон	—	23,08%	9,49%
Шум	4,78%	—	5,48%
Пауза	6,61%	13,37%	—
ИТОГО	11,39%	36,45%	14,97%

Таблица 3. Распределение ошибок классификации «тон/шум/пауза» по отдельным классам при обучении классификатора на материале речевого корпуса TIMIT, а тестировании — на материале речевого корпуса NTIMIT.

	Ошибка выделения тона	Ошибка выделения шума	Ошибка выделения паузы
Тон	—	24,04%	7,51%
Шум	1,56%	—	0,80%
Пауза	19,68%	59,05%	—
ИТОГО	21,24%	83,09%	8,31%

Средняя ошибка классификации «тон/шум/пауза» в первом эксперименте (обучение и тестирование на TIMIT) составила 16,43%, а во втором эксперименте (обучение на TIMIT, тестирование на NTIMIT) — 28,49%. В таблице 2 показано распределение ошибок классификации по каждому из классов «тон», «шум» и «пауза» для первого эксперимента, а в таблице 3 — для второго эксперимента.

Общая длительность звукозаписей, входящих в тестовое множество TIMIT, составляет 1 час 12 минут 16 секунд. Соответственно, такое же значение имеет и общая длительность звукозаписей, входящих в тестовое множество NTIMIT. Общая длительность вычислений в каждом из экспериментов составила 18 секунд для компьютера с двухъядерным ЦПУ Intel Core2Duo T2300 (тактовая частота ядра 1 ГГц) и объёмом ОЗУ 1,5 Гб. Таким образом, можно говорить о том, что нейросетевой алгоритм выделения тоновых, шумовых и паузных участков устной речи может работать как в реальном масштабе времени, так и в масштабе времени, ускоренном до 240 — 250 раз.

Выводы

Разработан нейросетевой алгоритм выделения тональных, шумовых и паузных участков устной речи. Проведены экспериментальные исследования на материале речевых корпусов TIMIT и NTIMIT, направленные на оценку точности, надёжности и скорости работы этого алгоритма.

В результате экспериментов оказалось, что разработанный алгоритм, работая в дикторонезависимом режиме, с высокой точностью выделяет тональные, шумовые и паузные участки устной речи, допуская лишь около 16% ошибок. Кроме того, алгоритм продемонстрировал высокую скорость работы: на базе аппаратного обеспечения типичного персонального компьютера выделение тоновых, шумовых и паузных участков устной речи было выполнено в ускоренном масштабе времени 240:1. Но при анализе речевых сигналов с нестационарным шумом, характерным для телефонного канала, количество ошибок алгоритма увеличилось приблизительно в 1,7 раза и составило около 28%. Наиболее типичным видом ошибки стало то, что тон (вокализованная речь) и шум (невокализованння речь) классифицировался как пауза (отсутствие речи). На наш взгляд, это связано с тем, что в условиях шума на паузных участках наблюдается достаточно высокий уровень энергии, сравнимый с уровнем энергии на участках речи.

Исходя из вышеописанного, можно сделать следующие выводы.

1. Разработанный алгоритм рекомендуется к использованию в системах распознавания речи, функционирующих в условиях отсутствия шумов или наличия стационарного шума. Такие условия характерны, например, для тихого офисного помещения или салона автомобиля.

2. Дальнейшие исследования будут направлены на повышение надёжности работы нейросетевого алгоритма в условиях нестационарного шума за счёт использования более инвариантной системы признаков (при этом система признаков по-прежнему должна оставаться простой и легко вычисляемой).

Литература

1. *Atal B.*, Rabiner L. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition // Acoustics, Speech and Signal Processing. – 1976. – Vol.24, Issue 3. – P.201-212.
2. *Jamal Ghasemi*, Amard Afzalian, M.R. Karami Mollaei. A Combined Voice Activity Detector Based On Singular Value Decomposition and Fourier Transform // Signal Processing. – 2010. – Vol.4, Issue 1. – P.54-61.
3. *Jankowski C.*, Kalyanswamy A., Basson S., Spitz J. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database//Proc. of ICASSP-90. – 1990. – P. 109-112.
4. *LeCun Y.,* Bottou L., Orr G., Muller K. Efficient BackProp // Neural Networks: Tricks of the trade. – Springer Verlag, 1998. – P. 5-50.
5. *Martin A.,* Charlet D., Mauuary L. Robust speech/non-speech detection using LDA applied to MFCC // Proc. of ICASSP'01. – 2001. – Vol.1. – P.237-240.
6. *Wilson D.R.*, Martinez T.R. The general inefficiency of batch training for gradient descent learning // Neural Networks. – 2003. – Vol.16. Issue 10. – P.1429-1451.
7. *Zue V.*, Seneff S., Glass J. Speech database development at MIT: TIMIT and beyond // Speech Communication. – 1990. – Vol. 9, № 4. – P.351-356.
8. *Архипов И.А.*, Гитлин В.Б., Лузин Д.А. Адаптивный алгоритм принятия решения «ТОН — НЕ ТОН», синхронный с основным тоном // Речевые технологии. – 2009. – №1. – С.80-93.
9. *Горбань А.Н.* Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей // Сибирский журнал вычислительной математики. – 1998. – Т.1, № 1. – С. 12-24.
10. Методы обработки речевых сигналов во временной области / Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. Пер. с англ. – М.: Радио и связь, 1981. – С.110-160.
11. Однонаправленные многослойные сети сигмоидального типа / Осовский С. Нейронные сети для обработки информации. Пер. с польск. – М.: Финансы и статистика, 2004. – С.46-88.

УДК 004.394:534.78

Нейромережевий алгоритм виділення тональних, шумових і паузних ділянок устного мовлення

І.Ю. Бондаренко[1], **О.М. Ладошко**[2]

[1] Донецький національний технічний університет,

в ул. Артема, 58, Донецьк, 83000, Україна

[2] Національний технічний університет України «КПІ»,

пр. Перемоги, 37, Київ-56, 03056, Україна

Розглядається проблема автоматичного виділення тональних, шумових і паузних ділянок усного мовлення. Для вирішення цієї проблеми пропонується нейромережевий алгоритм, що виконує класифікацію послідовності фреймів, на які розбивається мовний сигнал. На матеріалі мовних корпусів TIMIT і NTIMIT проведені експерименти оцінкм якості, надійності і швидкості роботи алгоритму в дикторонезалежному режимі, у тому числі в умовах нестаціонарного шуму, викликаного впливом телефонного каналу. Библ. 11, рис. 2, табл. 3.

Ключові слова: *класифікація мовлення-шум-пауза, нейронна мережа.*

Neural network algorithm for detection tonal, noise and pauses parts of continuous speech

I.U. Bondarenko[1], **O.N. Ladoshko**[2]

[1] Donetsk National Technical University

58 Artema, Donetsk 83000, Ukraine

[2] National Technical University of Ukraine «KPI»,

37 Prospect Peremogy, Kiev 03056, Ukraine

The problem of automatic detection of tones, noise and pauses parts of speech is considered. To solve this problem, we propose a neural network algorithm to classify sequences of frames into which the speech signal is separated. On the material of speech of corpuses TIMIT and NTIMIT experiments on evaluation of the quality, reliability and speed of the algorithm in speaker independent mode, including in non-stationary noise caused by the influence of the telephone channel were implemented. Reference 11, figures 2, table 3.

Key words: *classification of speech-to-noise-pauses, the neural network.*

1. *Atal B.*, Rabiner L. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition // Acoustics, Speech and Signal Processing. – 1976. – Vol.24, Issue 3. – P.201-212.
2. *Jamal Ghasemi*, Amard Afzalian, M.R. Karami Mollaei. A Combined Voice Activity Detector Based On Singular Value Decomposition and Fourier Transform // Signal Processing. – 2010. – Vol.4, Issue 1. – P.54-61.
3. *Jankowski C.*, Kalyanswamy A., Basson S., Spitz J. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database//Proc. of ICASSP-90. – 1990. – P. 109-112.
4. *LeCun Y.*, Bottou L., Orr G., Muller K. Efficient BackProp // Neural Networks: Tricks of the trade. – Springer Verlag, 1998. – P. 5-50.
5. *Martin A.*, Charlet D., Mauuary L. Robust speech/non-speech detection using LDA applied to MFCC // Proc. of ICASSP'01. – 2001. – Vol.1. – P.237-240.
6. *Wilson D.R.*, Martinez T.R. The general inefficiency of batch training for gradient descent learning // Neural Networks. – 2003. – Vol.16. Issue 10. – P.1429-1451.

7. *Zue V.*, Seneff S., Glass J. Speech database development at MIT: TIMIT and beyond // Speech Communication. – 1990. – Vol. 9, № 4. – P.351-356.
8. *Arkhipov I.A.*, Gitlin V.B, Luzin D.A An adaptive algorithm for deciding "TONE - not tone", synchronous with the main tone / / Speech technologies. – 2009. - № 1. - P.80-93. (Rus)
9. *Gorban A.N.* Generalized approximation theorem and the computational capabilities of neural networks / / Siberian Journal of Numerical Mathematics. –- 1998. - V.1, № 1. - P. 12-24. (Rus)
10. *Methods* of processing speech signals in the time domain / L.R. Rabiner, R. Schafer Digital processing of speech signals. Per. from English. - M.: Radio and communication, 1981. – P.110-160. (Rus)
11. *Unidirectional* multilayered network of sigmoidal type / Osovsky C. Neural networks for information processing. Per. from Polish. - Moscow: Finance and Statistics, 2004. - P.46-88. (Rus)

Поступила в редакцию 13 ноября 2012 г.