UDC: 004.93

A method of reducing the search for words in speech recognition based on phonetic coding algorithm

J.M. Du[1] and V.S. Vykhovanets[2]

[1]Bauman Moscow State Technical University, Moscow, Russia
[2]Institute of Control Sciences RAS, Moscow, Russia

forsola.du@gmail.com, valery@vykhovanets.ru

Abstract

This paper considers the problem of reducing the search for words in speech recognition of single word in a large Russian dictionary. Experimental data show the comparison of the recognition effectiveness and accuracy between the classic system and a modified one. The paper presents one potential direction for speeding up the HMM-based recognition process by using phonetic coding algorithm like SoundEx.

Keywords: speech recognition, hidden Markov model, phonetic algorithm, phonetic coding, allophone, acceleration of recognition

1. Introduction

Modern automatic speech recognition systems are usually HMM-based[1]. Due to the large-scale statistical data, the accuracy is becoming higher. However, as data continues growing, dictionary is also needed to be updated, and the decoding process takes more time as well. There are many researches [2, 3] aim at this problem, however there are little researches focus on the efficiency of word search in the large dictionary. Many technical companies, such as Microsoft, Google and Amazon, are providing their recognition services. Because of the problem of the huge amount of calculations, all these service systems need huge sever networks for reducing the calculation time.

Although most of these services are not open sources, according to the official researches like [10], the basic structure of some systems has been presented, in which a HMM based decoder [11] is described in the system. So, the HMM-based open source platform can be used for the approximate parallel analysis and comparison.

In this paper, a method based on the idea of phonetic algorithms [4] would be discussed by using the CMUSphinx recognition toolkit to recognize single word.

This method reduces the volume of dictionaries used in the decoding process via preprocessing, which is based on acoustic feature, to release the burden of search. By doing so, it can reduce the time of decoding in recognizing single word. Subsequently, the accuracy and efficiency of the method would be examined by experiments. The result shows this method can speed up the HMM-based system.

2. Principle of phonetic coding algorithm

The most famous phonetic coding algorithm is SoundEx. This algorithm applies a coding method designed to eliminate spelling and typographical errors in names. This algorithm finishes coding process according to pronunciation of words, and then based on phoneme — the smallest part of speech. The basic principle of letter coding in the algorithm is that consonants and vowels in a word are integrated and classified by numbers, and letters that are similar in sound are encoded with the same number, unpronounceable letters are deleted, and the first pronounceable letter is remained as it was [4].

3. Architecture of new recognition system

Many research institutions have implemented their own hidden Markov model (HMM) toolkit, such as Kaldi [7], CMUSphinx [6], and HTK [5]. CMUSphinx is selected as an experimental tool because its codes, typical recognition structures are more mature and open. Moreover, it has a ready-made acoustic model and a Russian language model with a large dictionary (540 thousand word).

In the recognition process of CMUSphinx, the decoder will use these 3 files to compute — acoustic model, pronunciation dictionary and language model. Pronunciation dictionary, or phonetic dictionary, contains all words and it presents their possible variants of pronunciation by phonemes. The decoder will scan all possible sequences of words from dictionaries to map the best path in the next process.

Therefore, it is planned to add one operation — "phonetic processing" before decoding. The new architecture of system is shown in Fig. 1. This part applies simple acoustic features of words and separates one abridged dictionary with words that have similar pronunciation with these features. It will achieve the goal of reducing decoding time in the process of recognition of single word. The simplest and most famous algorithm for this operation is phonetic algorithm.

4. Theory of the phonetic coding

CMUSphinx provides a function called "Allophone". Only by using the acoustic model and simple rules for segmenting phonemes can this function outputs a sequence of phonemes without using dictionary and language models. This function only needs very short time. The examples of outputs are shown in Tab. 1.

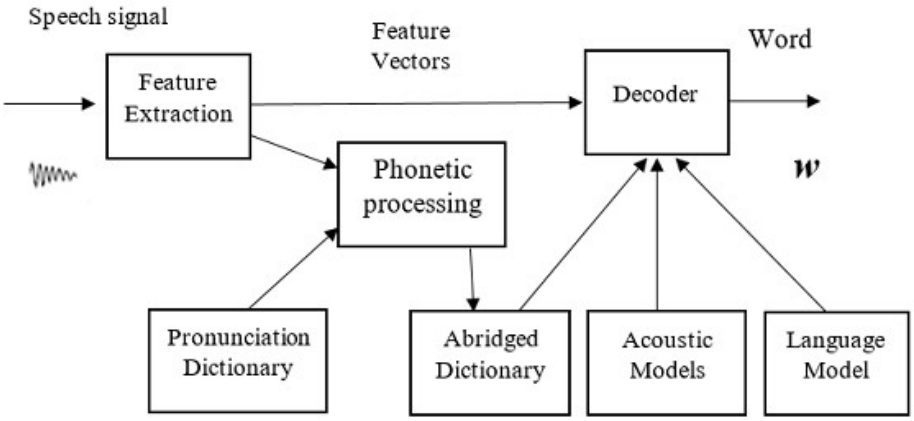

Fig. 1. The architecture of new recognition system

Word	Transcription	Allophones
дедушка	dj e1 d u0 sh k a0	bj dj e0 d u0 sh kj a0
ресница	rj e0 s nj i1 c a0	e0 z n1dj z a1 m
поплакать	p o0 p l a1 k a0 tj	k a0 p l a1 g j a0

Table 1. Examples of outputs from the function "Allophone"

From examples, although the relevant results do not completely coincide with the transcriptions of words from the dictionary, their "similarity" can be seen. It is no need to look through the entire dictionary for word searching, part of it that contains "similar" words is required while another part with those "dissimilar" ones in the dictionary is ignored. In other words, if we can find the "rules of the similarity" between transcription and result of allophone and use these to do a preliminary search and extract words for the dictionary, we can create an "abridged dictionary" for the next recognition process.

In experiments, 8-hour audio material consisting of 7000 Russian words spoken by 4 announcers (2 men and 2 women) was applied. Among the 7000 words, 5000 words are allocated for the finding of rules for constructing the method. In addition, the remained 2000 word are used to analyze its effectiveness.

After an artificial analysis, there are 3 phonetic rules for allophones that can be used to form the abridged dictionary.

4.1. The length of the sequence of allophones and the length of the transcription of the words. From the example in Tab. 1, their sequence lengths are very similar. Although they are not completely equal, we can introduce a range of errors for the length of the sequence. Therefore, after adding an error range to

the length of the allophones, the length of the transcription meets the condition of range. Subsequently, it can be assumed that these lengths for this range are "equal".

Tab. 2. shows the range of errors, the coverage and the amount of words, which are fallen within each range.

Length of allophones	Range of errors	Coverage	Amount of words in the range
1	[1, 4]	100.0%	616
...	...	...	...
8	[4, 11]	98.4%	4504
...	...	...	...
12+	[9, 18+]	100%	1263

Table 2. The ranges of errors for lengths of allophones.

The coverage is defined as below.

For length of allophones N, there are X words, whose length of allophones is N. In these words, Y words have the length of transcription, which are "equal" to N for the range of errors, then the coverage of allophones N is defined as following,

$$Coverage_N = Y/X. \tag{1}$$

This coverage shows the coverage of the range of errors. As its height has to be ensured, there are less correct words out of the coverage. The amount of words in each range shows the uniqueness of each range. The amount of words in each range has to be decreased, so that a smaller dictionary can be built for each range. Therefore, a larger range is determined. The higher coverage is, the bigger the amount of words in the range is. It aims to minimize the amount of words in the range in the precondition ensuring correct recognition.

4.2. The number of vowels in the sequence of allophones and the relevant number in the sequence of transcriptions. Like the lengths of sequence, the numbers of vowels in both sequences are very close to each other. According to artificial statistics, the following ranges of errors for number of vowels are shown in Tab. 3.

4.3. The first consonant of a word. The first consonant is one of the most important features of a word. Due to reading errors, inaccurate acoustic model, imperfect speech segmentation and other reasons, the recognition of the first consonant is always accompanied by three errors.

a) Superfluous first consonant.

Numbers of vowels	Range of errors	Coverage	Amount of words in the range
1	[0, 2]	99.0%	2172
2	[1, 3]	94.6%	3741
3	[2, 4]	95.7%	4129
...	...	...	...
9+	[8, 10+]	100%	6

Table 3. The ranges of errors for number of vowels

The first consonant in transcription is not the first consonant in allophone sequence, but with other independent consonants existing before it. For example, the word "дедушка" in Tab. 1. It is established by artificial statistic that the similarity of the first consonant in transcription and allophone sequence is 69.7%. If it is assumed that the first consonant of transcription appears in the first few consonant allophones, the coverage will increase.

b) Absence of the first consonant.

The first consonant in transcription is absent in the sequence of allophones. For example, the word "ресница" in Tab. 1. It can be seen from statistical data that such error often arises when a vowel appears behind the consonant. Therefore, when the first phoneme of allophone sequence is a vowel, it is needed to consider the possibility of the absence of the first consonant.

c) Other first consonant.

For example, in Tab. 1, the first consonant of transcription "поплакать" is signed as "p"; in its result of allophone, it is incorrectly recognized as "k". However, this is not an accidental situation. The first consonant "p" in many words is recognized as "k". Therefore, it is needed to agree that "p" can be recognized as "k". It is not incorrect.

According to cases a), b), c) and the research [8, 9], the consonants is regrouped as Tab. 4.

4.4. Coding with the rules. From these rules, all the transcriptions and results of allophones just like this example can be coded as following.

For the word "дедушка", its transcriptions are "dj e1 d u0 sh k a0"; while relevant allophone results are "bj dj e0 d u0 sh kj a0".

For transcription, length of sequence – 7, number of vowels – 3, first consonant is in group 3, with code as "7-3-3".

For allophones, length of sequence – 8, number of vowels – 3, first consonant is in group 4, second consonant is in group 3, with code as "8-3-4-3".

Group	First phoneme of allophones	Possible first consonant of transcription
1	k, kj	b, bj, g, gj, k, kj, p, pj
2	p, pj	b, bj, k, kj, p, pj, t, tj
3	d, dj	d, dj, k, kj, z, zh, zj
4	b, bj	b, bj, d, dj, p, pj
...	...	...
9	vowels	b, bj, g, gj, k, kj, m, mj, n, nj, p, pj, t, tj, z, zj, zh

Table 4. Group of the first phoneme of allophones and possible first consonant of transcription

After coding, 2 kinds of code are got, the "code of transcriptions" and "the code of allophones". These codes will not be directly compared. It is necessary to translate from the code of allophones to that of transcriptions. The steps of translation are as follows.

1) The code of transcriptions is used for all words in the dictionary. Therefore, a full dictionary with codes of transcriptions will be obtained.

2) The code of allophones is used for defining similar words. The code of allophones and their ranges from Tab. 2, Tab. 3 and Tab. 4, would be applied to select several codes that satisfy all ranges – supposing that the words with these codes are similar.

3) The words would be rewritten from step 2) into a new dictionary. Subsequently, an abridged dictionary with less volume is got.

Tab. 5 shows the translation result of code '8-3-4-3'. From the table, it can be found that the code '7-3-3' is included in the results.

4-2-4	5-2-4	6-2-4	7-2-4	...	11-2-4
4-3-4	5-3-4	6-3-4	7-3-4	...	11-3-4
4-4-4	5-4-4	6-4-4	7-4-4	...	11-4-4
4-2-3	5-2-3	6-2-3	7-2-3	...	11-2-3
4-3-3	5-3-3	6-3-3	**7-3-3**	...	11-3-3
4-4-3	5-4-3	6-4-3	7-4-3	...	11-4-3

Table 5. Translation result of the code of allophones "8-3-4-3"

Fig. 2 shows components of the "Phonetic processing" in Fig. 1.

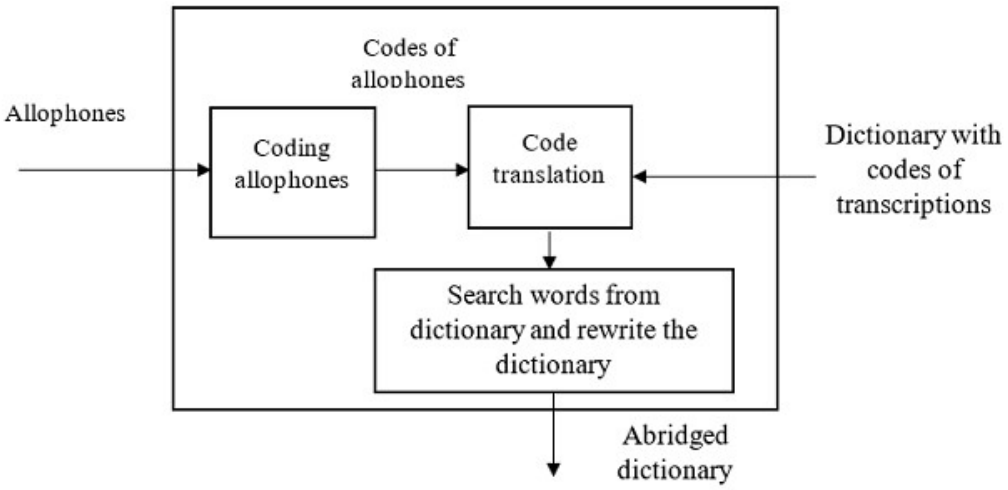

Fig. 2. Components of the "Phonetic processing"

5. Experiment

Experiments are conducted from 2 aspects — the effectiveness and the correctness of the system.

In the first experiment of recognizing 2000 words by using a classic recognition system and the modified system, the relevant result of the average time of recognizing single words are shown in Tab. 6.

Process	Classic system (second)		Modified system (second)	
	Confidence interval at 0.95	Average time	Confidence interval at 0.95	Average time
Function "Allophone"			[0.34, 0.51],	0.044
Phonetic processing			[0.039, 0.049]	0.42
HMM based decoder	[2.34, 3.46]	2.90	[0.73, 1.09]	0.91
Total time	[2.34, 3.46]	2.90	[1.21, 1.55]	1.38

Table 6. Results of "effectiveness experiments"

From Tab. 6, the modified system is noticeably faster than the classic one.

According to the official researches like [10], the basic structure consisted of a HMM based decoder [11] and a neural networks based acoustic model. This whole structure is very similar as the one of another recognition platform Kaldi [7]. So,

we well use Kaldi as an approximate parallel comparison of the acceleration in the HMM-based system.

The standard metric used to quantify the speed of an speech recognition system is the real-time factor (xRT), defined as the ratio between the time it takes to process the input and the duration of the input. In Tab. 7 we comparative the real-time factor of the Sphinx classic system, the Kaldi classic system and our modified Sphinx system with different volume dictionary. Need to mention, because of the lack of existing Russian data in Kaldi, we use the English data for its experiment.

Volume of dictionary	Sphinx classic (xRT)	Kaldi classic (xRT)	Sphinx modified (xRT)
7k	1.46	1.28	1.57
40k	2.12	3.24	1.85
500k	5.80	8.12	2.61

Table 7. Comparison of real-time factors

The result shows, that the method can speed up the HMM-based recognition when large dictionary is used.

In "accuracy experiments", the relevant results of the accuracy of the 2 systems are shown in Tab. 8.

	Classic system	Modified system
Accuracy of recognition of single words	56.1%	61.7%

Table 8. Results of "accuracy experiments"

The accuracy of the modified system is higher than that of the classic one. From an artificial check of the relevant results, it can be found that many words in the classic system are recognized as word combinations (phrases) with short words. The main reason for this phenomenon is that these phrases have more probability in language model. In addition, the modified system uses the rules to determine the length of a word in a certain range. Therefore, in its dictionary, there is no too-short words due to the improvement of accuracy.

6. Conclusion

In this article, one method based on the idea of the phonetic coding algorithm and the rules of allophones is described. The method shortens the search of words

from the dictionary in the recognition process of a single word in a large dictionary. Corresponding experimental results show that this method increases the average speed of the decoder process by about 52% (average time from 2.9 second to 1.38 second). In addition, its accuracy is higher than that of the classic system (from 56.1% to 61.7%). Thus, this article represents one possible direction for speeding up the recognition process.

REFERENCES

1. Gales M., Young S. The Application of Hidden Markov Models in Speech Recognition //Foundations and Trends in Signal Processing. 2007. V. 1. No. 3. P. 195-304.
2. Mosle M. Accelerating Speech Recognition Algorithm with Synergic Hidden Markov Model and Genetic Algorithm Based on Cellular Automata // International Conference on Signal Processing Systems. Singapore. May 15-17, 2009. P. 3-8.
3. Yu L., Ukdave Y. GPU-accelerated HMM for Speech Recognition // Parallel Processing Workshops. Minneapolis USA. September 9-12, 2014. P. 395-403.
4. Vykhovanets V., Du J.,Sakulin S. Overview of phonetic encoding algorithms// Large-Scale Systems Control. Issue 73, Moscow: Institute of Control Sciences RAS. 2018. P. 67-94.
5. Aiman F., Saquib Z., Nema S. Hidden Markov Model system training using HTK//International Conference on Advanced Communication Control and Computing Technologies. Ramanathapuram, India. May 25-27, 2016. P. 806-809.
6. Lamere P., Kwok P. The CMU Sphinx-4 speech recognition system// IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China. April 6-10, 2003.
7. Povey D., Ghoshal A. The KALDI Speech Recognition Toolkit// IEEE Workshop on Automatic Speech Recognition and Understanding. Hawaii, USA. December 11-15, 2011.
8. Paramonov V.V., Shigarov A.O. Polyphon: An Algorithm for Phonetic String Matching in Russian // Language International Conference on Information and Software Technologies. Druskininkai, Lithuania. October 13-15, 2016. P.568-579.
9. Skripnik Y.N., Smolenskaya T. M. Phonetics of the modern Russian language: a textbook. SGPI, Stavropol. 2010.
10. Xiong W., Droppo J. The Microsoft 2016 Conversational Speech Recognition System// IEEE International Conference on Acoustic, Speech and Signal Processing. New Orleans, USA. March 5-7, 2017. P. 5255-5259.

11. Mendis C., Droppo J Parallelizing WFST speech decoders// IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China. March 20-25, 2016. P. 5325-5329.

УДК: 004.5

Извлечение знаний из больших объёмов данных по шаблонам логики тайлов

В.В. Девятков[1]

[1]МГТУ им. Н.Э. Баумана, 2-я Бауманская ул., 5, Москва, Российская Федерация

[2]ИПУ РАН, Профсоюзная ул., 65, Российская Федерация

deviatkov@bmstu.ru, deviatkov@comtv.ru

Аннотация

Настоящая статья посвящена изложению похода к извлечению знаний о поведении распределённых взаимодействующих объектов путём анализа закономерностей их поведения на основе частных шаблонов поведения по общим данным взаимодействия, извлекаемым из больших объёмов данных в каналах связи. В качестве языка для решения задач извлечения знаний и описания шаблонов выбрана логика тайлов. В статье обосновывается такой выбор, рассматриваются принципы формирования шаблонов в логике тайлов и сущность их дальнейшего использования. Принципы формирования и использования шаблонов для извлечения знаний иллюстрируется примером анализа поведения простейшей клиент-серверной системы. Обсуждаются перспективы использования предлагаемой методики для различных приложений.

Ключевые слова: Искусственный интеллект, извлечение знаний, большие данные, шаблоны, логика тайлов.

1. Введение

В связи со все ускоряющимся прогрессом в области искусственного интеллекта, одной из востребованных задач в этой области стала задача анализа поведения различных взаимодействующих в сети объектов: агентов, клиентов, серверов, разного рода пользователей и т.п. Этот анализ приходится осуществлять путём наблюдения за большими потоками информации (данных) указанного взаимодействия. При этом архитектура системы взаимодействующих объектов, как правило, считается известной и задачей анализа является выявление каких-либо закономерностей взаимодействия Настоящая работа основана

Работа выполнена при финансовой поддержке Минобрнауки, проект №2.5048.2017/8.9 от 01.01.2017.