UDC 519.688:004.032.26

Informatics and Mathematical Methods in Simulation
Vol. 6 (2016), No. 3, pp. 297-301

DEVELOPMENT OF EFFECTIVE VOCABULARY STRUCTURES FOR THE SPEECH RECOGNITION TASKS

D.V. Zahanich, I.E. Mazurok

Odessa I.I. Mechnikov National University,
Dvoryanskaya st., 2, Odessa, 65026, Ukraine; e-mail: igor@mazurok.com

In this paper we describe a speech recognition method, which is optimized for mobile devices with limited computing power. This article is focused on reducing the size of a necessary dictionary and development of method of finding dictionary strings that match the input speech data. The proposed approach allows to solve the problem of speech recognition on mobile devices offline.

Keywords: speech recognition system, dictionary structure, fuzzy search.

Introduction

In recent decades mankind faced the problem of creating effective tools for data input. Keyboards and touch screens don't solve the problem efficient enough, because these tools are not natural for humans. Today personal computers are used not only by specialists, who are capable of fast typing, but also by regular users with poor computer skills. To solve this problem, various techniques and special devices for rapid data entry were developed. For example, sign writing received a new life and many devices allow you to enter words with line drawing on the touch screen. However, the core problem remains the same, because such methods also require special skills.

There are cases in which a user can't perform data input with regular tools or it is very difficult to do so. For example, it is quite difficult to type a message while driving a car. Moreover, the number of such scenarios increases with the spread of portable devices such as smartphones, automobile navigation devices, "smart home" systems, etc.

Conventional data input methods require performing some actions by user’s hands and often prohibits user from performing others tasks at the moment of data input. However, speech recognition methods don’t have these disadvantages. It took almost half a century to improve speech recognition systems for widespread usage [1]. Researchers and developers of such systems had to solve a lot of problems [2]:

- the problem of the extraction of the desired information from an audio signal;
- the problem of the classification of the extracted information from an audio signal;
- necessity of extensive training of speech recognition systems because of a large number of words pronunciation variants;
- low overall performance of speech recognition systems.

In recent years, speech recognition systems have become much more popular than before. This trend started due to progress made in speech technology area and accumulation of large amounts of data in the Internet. However, modern speech models require a considerable memory capacity and computation power. In this regard, the modern speech recognition systems perform most of calculations on the server side with multi-core processors and CPU/GPU clusters with great computational power [3]. Thus, the problem of low performance of speech recognition systems has not been solved yet. This becomes the major problem in case of mobile devices that can't be constantly connected to the Internet.

The primary *goal* of this article is to develop speech recognition method, which is capable of running on the mobile device with limited computational power. We focus on reducing the size of the dictionary, which is the key component of every speech recognition system.

Article

A proposed method of speech recognition is based on fuzzy string search. Thus, the speech recognition is achieved by searching the recognized phonetic transcription of the spoken word in the phonetic dictionary.

Speech sounds are infinitely varied. Accurate physical analysis may reveal that a person never produces exactly the same sounds. For example, the “k” sound in the word “kit” and “k” sound in word “skill” are not identical. However, these sounds are still represent the same phoneme /k/.

There are many sets of phonemes that are used in speech recognition tasks. In this article we use the following set of phonemes of the English language, based on Arpabet [7]. Arpabet is phonetic transcription code developed by Advanced Research Projects Agency (ARPA) as part of their Speech Understanding Project (1971-1976). This code represents each phoneme in the American dialect of English language as a distinct sequence of ASCII characters. The current phoneme set contains 39 phonemes (or more accurately, "phones"). Basically, "phones" are more or less similar classes of sounds.

Table 1.

Phones of American dialect of English and their corresponding examples of pronunciation

Phoneme	Example	Phoneme	Example
AA	odd	L	lee
AE	at	M	me
AH	hut	N	knee
AO	ought	NG	ping
AW	cow	OW	oat
AY	hide	OY	toy
B	be	P	pee
CH	cheese	R	read
D	dee	S	sea
DH	thee	SH	she
EH	Ed	T	tea
ER	hurt	TH	theta
EY	ate	UH	hood
F	fee	UW	two
G	green	V	vee
HH	he	W	we
IH	it	Y	yield
IY	eat	Z	zee
JH	gee	ZH	seizure
K	key		

Example 1. Phonetic transcription of word “University” is represented in the form of code "Y UW N AH V ER S AH T IY".

To recognize a spoken word we acquire set of spoken phones and compare it with the phonetic transcriptions of the words, which are presented in the dictionary.

Human speech is a complex phenomenon. Variants of pronunciation of even the same word may be quite different from each other depending on a speaker, a speaker accent, level of external noise, and many other factors. In this regard, even completely correct set of recognized spoken phones will rarely be the same as conventional phonetic transcription of actually spoken word. To solve this problem, every word in the dictionary is represented as a set of possible pronunciations variants in form of phonetic transcriptions. Phonetic transcriptions of the words are encoded as ASCII characters, so we can use methods of fuzzy string search.

However, we must take into account the phenomenon of coarticulation. It leads us to the necessity of storing a large number of possible pronunciations of every word, which differ in similar-sounding phonemes.

To solve this problem we provide phonetic encryption algorithm based on Soundex.

In this modification of Soundex, similar-sounding phonemes are replaced to ASCII characters, and then every sequence of the similar characters is reduced to one character. The resulting line is called phonetic hash string.

Table 2.

Similar phonemes and their encoding characters

Character	Phonemes
A	AA, AO, HH, OW, AY, OY
B	B, P
F	F, TH
K	K
S	S, SH, Z, ZH
G	G, JH
V	V, W
T	D, DH, T, CH
L	L
N	M, N, NG
R	ER, R
E	EH, EY, AH, AE

Example 2. Pronunciations "Y UW N AH V ER S AH T IY" and "UW N EH V R S EH T IY" of word “University” are encoded into the same phonetic hash string "INEVRSETI".

The proposed phonetic encryption algorithm allows us to present all known pronunciations of the word in the dictionary as theirs phonetic hash strings, greatly reducing the size of dictionary.

The final stage of the proposed speech recognition method consists in searching an encoded phonetic string of spoken word in the dictionary. We use fuzzy string search algorithm, which is known as the FB-Trie algorithm (Eng. Forward-backward trie). It was described in the "Fast approximate search in large dictionaries" [4].

The task of fuzzy string search in the dictionary is to choose for a given search query W subset P of all the words from the dictionary D, which has distance p to the search query and p does not exceed a certain threshold N :

$$P = \{P_i \mid P_i \in D \bigcap p(P_i, W) \leq N\}.$$

We use the Damerau–Levenshtein distance because it is capable of detection up to 80% of all human misspellings.

The Damerau–Levenshtein distance is the distance between two strings, i.e., finite sequence of symbols, given by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters.

FB-Trie algorithm is based on the concept of universal Levenshtein automaton and it applies a dictionary structure in the form of direct and reverse prefix tree in conjunction with splitting of keyword W into two approximately equal parts W1 and W2. Reverse trie contains inversions of all the vocabulary words.

The choice of this algorithm is based on the comparative analysis in “Indexing methods for approximate dictionary searching” [5]. According to this research, the FB-Trie algorithm provides a linear search time depending on the length of the search query and dictionary structure consumes only 300% of the memory, which is occupied by the raw dictionary. Pronunciation dictionary of 20,000 words, which corresponds to the number of the most frequently used words in the English language, requires amount of memory that does not exceed 5-6 megabytes.

For comparison purposes, we take the statistics of RAM limit for applications on the Android mobile platform. According to the document "Android Application Memory Limit" [6], to support all possible device configurations speech recognition system should not consume more than 16 megabytes of RAM. Thus, the proposed method of speech recognition can be implemented even on the weakest of modern mobile devices.

Conclusion

This paper presents a method for speech recognition, which is based on fuzzy string search in pronunciation dictionary. The main advantage of this method is relatively small amount of memory, which is consumed by dictionary component, so it could be implemented on the devices with small computing power. The second advantage is simplified training process, which can be done even without recorded audio data. Efficiency and a small amount of memory required to build the software allow the implementation of the proposed algorithm for mobile communication devices and low-power computers. However, the proposed method doesn’t take into account context of speech and language grammar model, so it may show high error rate it continuous speech.

The proposed solution to speech recognition task can be used effectively in cases where the speech recognition system has limited computational resources and is not supposed to recognize complex grammatical structures. Such cases may include remote control systems with a small vocabulary of control commands.

References

1. Xuedong Huang. Spoken Language Processing: A Guide to Theory, Algorithm and System Development / Xuedong Huang, Alex Acero, Hsiao-Wuen Hon. – Prentice Hall PTR, 2001. – 480 p.
2. Zakaria Kurdi. Automatic Speech Processing and Natural Languages / Zakaria Kurdi // ISTE Wiley. Volume 1. – 2016. – 720 p.
3. Dong Yu. Automatic Speech Recognition: A Deep Learning Approach / Dong Yu, Li Deng. – Springer, 2014. – 328 p.
4. Mihov, S. Fast approximate search in large dictionaries / Stoyan Mihov, Klaus U. Schulz // Computational Linguistics. – 2004. – V.30. – No.4. - pp. 18-23.
5. Boytsov L. Indexing methods for approximate dictionary searching: Comparative analysis. / Leonid Boytsov // ACM J.Exp. Algor. 16. – 2011. – Vol. 1 – C. 47-65.
6. Android Application Memory Limit [Electronic resourse]. Available from: https://drive.google.com/file/d/0B7Vx1OvzrLa3Y0R0X1BZbUpicGc/view
7. Yukio Tono. Developmental and Crosslinguistic Perspectives in Learner Corpus Research / Yukio Tono, Yuji Kawaguchi, Makoto Minegishi. – John Benjamins Publishing, 2012. – 263 p.

РОЗРОБКА ЕФЕКТИВНИХ СТРУКТУР СЛОВНИКА ДЛЯ ЗАДАЧ РОЗПІЗНАВАННЯ МОВЛЕННЯ

Д.В. Заганич, І.Є. Мазурок

Одеський національний університет ім. І.І. Мечникова,
вул. Дворянська, 2, Одеса, 65026; e- mail: igor@mazurok.com

У статті пропонуються модифікації алгоритмів розпізнавання мови, оптимізованих для реалізації на мобільних пристроях з обмеженою обчислювальною потужністю. Головна увага приділяється скороченню розміру необхідного словника і алгоритму пошуку введених мовних даних. Запропонований підхід дозволяє вирішувати задачі розпізнавання мови на мобільних пристроях у автономному режимі.

Ключові слова: система розпізнавання мови, структура словника, нечіткий пошук

РАЗРАБОТКА ЭФФЕКТИВНОЙ СТРУКТУРЫ СЛОВАРЯ В ЗАДАЧЕ РАСПОЗНАВАНИЯ РЕЧИ

Д.В. Заганич, И.Е. Мазурок

Одесский национальный университет им. И.И. Мечникова,
ул. Дворянская, 2, Одесса, 65026; e-mail: igor@mazurok.com

В статье предлагаются модификации алгоритмов распознавания речи, оптимизированные для реализации на мобильных устройствах с ограниченной вычислительной мощностью. Основное внимание уделяется сокращению размера необходимого словаря и алгоритму поиска вводимых речевых данных. Предложенный подход позволяет решать задачи распознавания речи на мобильных устройствах в автономном режиме.

Ключевые слова: система распознавания речи, структура словаря, нечеткий поиск