THE ADVANTAGE OF INTERPHONEME PROCESSING AT DIPHONE RECOGNITION OF KAZAKH WORDS

Aigerim Buribayeva Altynbek Sharipbay
L.N. Gumilyov Eurasian National University
buribayeva@mail.tu; sharalt@mail.ru;

ABSTRACT

This paper presents a method of interphoneme processing at diphone recognition of Kazakh words. Authors made experiment to test how impact interphoneme processing to recognition accuracy. The experiment results show that recognizable word best differs from the other word on the DTW-Distance after interphoneme processing than without it. The results can be used in the construction of recognition system of single words.

1 Introduction

Automatic recognition of natural language verbal speech is one of important areas of development of artificial intelligence and computer science as a whole, as results in this area will allow to solve the problem of development of man’s efficient voice response means with the help of computer. A principal opportunity for transition from formal languages-mediators between man and machine to natural language in verbal form as universal means of expression of man’s ideas and wishes has appeared with development of modern voice technologies. Voice input has a number of advantages such as naturalness, promptness, input’s notional accuracy, user’s hands and vision freedom, possibility of control and processing in extreme conditions.

Specialists from several scientific areas research the problem of speech recognition for more than 50 years. Methods and algorithms which are used are separated into four big classes:

- Methods of discriminant analysis based on Bayesian Discrimination [1];
- Hidden Markov Model [2];
- Artificial neural networks [3];
- Dynamic programming - dynamic time warping (DTW) [4];

It should be noted a number of benefits sought by the development of speech recognition systems:

- Continuous speech - feature that allows users to speak naturally (continuous), not pausing between words (discrete speech input).
- Large dictionaries - the ability to process a large Word Count general and special categories of technical and subject areas of knowledge to increase the capacity and effectiveness of voice recognition systems.
- Independence from the speaker - the system's ability to recognize words without personal computer settings by repeating the same speech.

The most frequently and successfully for recognition of continuous speech using Hidden Markov Model (HMM) [5, 6] or Artificial Neural Networks [6, 7]. Different base units: phonemes, allophones, diphones and triphones, etc. selected for speech recognition. Dynamic time algorithms (DTW) still effective to recognize single words [8].

We chose the word recognition technology based on the collected diphone database, because single words recognize is more accurately [9]. The system does not recognize the diphones separately, it synthesizes of these the words' etalons, and then recognize whole words by the algorithm DTW. The advantage of the system is that to add a new word there is no need to train the system voicing the word, but rather enter a word in text form. Automatic generation of words' etalons of diphones will make a step towards large dictionaries, and speaker-independent systems can be achieved by averaging the etalons.

2 Materials and Methods

According to [10], the authors have decided to formulate its present viewpoint in the following thesis: «One of the possible keys to speech recognition lies in interphoneme transitions.»

Analysis of the situation, we can start with the following simple experiment. Using any known program for working with sound, for example «Sound Forge», write any two words, and then cut out, fixed (middle) part of their component sounds. Reproducing the resulting audio signals, we can at the hearing to determine what the words sound. On the contrary, by cutting interphoneme transitions, and leaving the stationary part of the sound, we found it difficult to distinguish by ear, for example, words «шана» и «сана».

So, it was a program which allows using the diphone database, automatically generate etalons given dictionary of words and keep them DTW-recognition. we have described in detail the construction of such a system in [10]. Etalons of words recognized by the dictionary formed from the etalons of diphones, which database of approximately one and a half thousand created for each speaker in advance [9]. Creation of such a database in the future eliminates the need to create any etalons by voice. we mean that the corresponding diphone interphoneme transition within a word, the site in standard lengths: 3 windows in the 368 samples to the left of the label between the sounds and 3 of the same window to the right of the same label. Etalon of diphone - set of 6 appropriate vectors. In addition, we use a section of 3 windows at the beginning of words and site in 3 windows at the end of words, conditionally call them respectively the initial and final middiphone speech (the transition from silence to speech and vice versa). All vectors in etalons diphones, play the role of the code vectors and form a codebook B. All etalons diphones are numbered, numbered and all the code vectors. Every word of the dictionary automatically transcribed, transcription construct string names diphones. Each of them is replaced by the corresponding diphone's etalon. The resulting string vectors forms a word's etalons.

We apply for recognition has already become a classic algorithm T. Vintsyuk known as algorithm DTW. We use the feature vector related to the relative frequencies of the lengths of complete oscillations in the speech segments in 368 samples [9].

The recognition process is constructed as follows. Recognize words automatically segmented and then subjected to interphoneme processing: removed stationary components of the sounds and left only diphones around labels between sounds (interphoneme transitions). And only then word gets recognition.

It is known that the DTW-word recognition with etalons built of diphones, perhaps as a signal in which the stationary part of the deleted sounds (interphoneme processing) and for the original signal.

In this regard, we decided to check how interphoneme processing affects to recognition accuracy.

3 Experiments

We made experiment to test how impact interphoneme processing to recognition accuracy. We chose 100 of Kazakh word for recognition. Only one female speaker participated in the experiment because our system recognizes a specific speaker. First, we recognize the words in the mode "No interphoneme processing." After the same words were recognized in the mode "with interphoneme processing". The experiment was done in a regular university’s classroom without noise isolation.

4 Results

The results of the experiment are shown in the following table:

Table 1 - Result of recognition of the word "Қазақ" ("Kazakh")

Word	**DTW-distance without** *interphoneme processing*	**DTW-distance with** *interphoneme processing*
Қазақ (Kazakh)	15,88	10,32
Намыс (pride)	23,84	20,74
Бетеге (fescue)	30,77	26,35
Өнер (talent)	25,57	25,55
Араша (pull apart)	24,83	23,37

The table shows the result of the recognition of the word "Kazakh". The second column shows DTW-distance of 5 words at the most immediate recognition in the "No interphoneme processing." In the third column shows DTW - distance of 5 words at the most immediate recognition in the "interphoneme with treatment."

As you can see, in the first case relation of two first distances in column is:

$$\frac{23{,}84}{15{,}88} \approx 1{,}50 .$$

In the second case it is equal to

$$\frac{20{,}74}{10{,}32} \approx 2{,}00 .$$

Analogous result is visible in all our other experiments.

Recognizable word best differs from the other word on the DTW-Distance after interphoneme processing than without it. Conclusion: the recorded speech signal is advisable to expose interphoneme at diphone recognition.

5 Acknowledgments

The presented work is supported by “Automation of Recognition and Generation of the Kazakh Language Written and Oral Speech” Project implemented under the budget program 120 “Grant Financing of Scientific Researches”, specificity 149 “Other Services and Works”, by Priority 3. Information and Telecommunication Technologies.

6 References

[1] Raut, C.K., Bayesian discriminative adaptation for speech recognition, Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on Eng. Dept., Cambridge Univ., Cambridge , 19-24 April 2009, Page(s): 4361 – 4364

[2] Lawrence, R. Rabiner (February 1989). "A tutorial on Hidden Markov Models and selected applications in speech recognition". Proceedings of

the IEEE 77 (2): 257–286. doi:10.1109/5.18626.

[3] Al-Alaoui, M.A., Al-Kanj, L., Azar, J., and Yaacoub, E., Speech Recognition using Artificial Neural Networks and Hidden Markov Models, IEEE MULTIDISCIPLINARY ENGINEERING EDUCATION MAGAZINE, VOL. 3, NO. 3, SEPTEMBER 2008

[4] Винцюк, Т.К. Анализ, распознавание и интерпретация речевых сигналов. Киев, Наук. думка, **1987.**

[5] Najkar, N., Razzazi, F., Sameti, H. An evolutionary decoding method for HMM-based continuous speech recognition systems using particle swarm optimizationб Pattern Anal Applic, DOI 10.1007/s10044-012-0313-7

[6] Frikha, M., Ben Hamida, A. A Comparitive Survey of ANN and Hybrid HMM/ANN Architectures for Robust Speech Recognition American Journal of Intelligent Systems 2012□ 2(1): 1-8 DOI: 10.5923/j.ajis.20120201.01

[7] Hosom, J.P., Cole, R., and Fanty, M. Speech Recognition Using Neural Networks at the Center for Spoken Language Understanding. //Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, July 1999.

[8] Dev Dhingra, S., Nijhawan, G., Pandit, P., Isolated Digit Recognition Using MFCC AND DTW, International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), ISSN (Print): 2278-8948, Volume-1, Issue-1, 2012, pp 59-64

[9] Шелепов, В.Ю., Ниценко А., Дорохина, Г.В., Карабалаева, М.Х., Бурибаева, А.К. О распознавании речи на основе межфонемных переходов. Вестник. Астана: Евразийский национальный университет им. Л.Н.Гумилева, 2012. – Специальный выпуск.–С.436-440

[10] Бурибаева, А.К. Распознавание казахских слов на основе дифонной базы, Труды Международной конференции "Компьютерная обработка тюркских языков" (Turklang-2013), С. 230-239.