Isolated Speech Word Recognition Based on Fuzzy Pattern Matching with Optimal Temporal Alignment

Oleg I. Fedyaev [1], *Ivan Yu. Bondarenko* [1]

[1] Department of Applied Mathematics and Computer Science
Donetsk National Technical University, Ukraine
fedyaev@r5.dgtu.donetsk.ua, bond005@yandex.ru

Abstract

Fuzzy Pattern Matching application is considered for automatic isolated word recognition. An Optimal Temporal Alignment algorithm is proposed for temporal instability problem solution of speech patterns under comparison based upon the maximal similarity criterion. The results of experimental research in single-speaker and multi-speaker recognition are described for a 105 words vocabulary using the algorithm proposed.

1. Introduction

Speech commands utilization for technologic systems' control, especially robots, is getting more and more required by users of such systems. Audio speech as the medium of information exchange is most natural for people. A user needs no special training to operate the audio speech at human-machine interaction. Besides, it often happens that audio speech is the unique form of dialogue for people suffering from malfunctions of locomotorium and vision. To apply a speech control channel within the environment of intensive information interchange between an operator and an operated device means to free his eyes and hands thus allowing the full attention concentration on control procedures [1]. The problem solution of speech recognition in speech-controlled systems must satisfy the following requirements:

- high-precision recognition of the limited number of isolated speech words i.e. control commands;
- the system functioning in speaker-dependent and speaker-independent manner.

The basic methods for the problem solution are based on probabilistic, metrical and neural network approaches. The approach based on the fuzzy logic [2] can be prospective in the problems solution characterized by difficult formalization including the above mentioned problem of isolated speech word recognition.

A number of methods in known using the fuzzy logic for recognition of isolated words. So, the paper [3] offers a hybrid system based on the principles' combination of dynamic programming and fuzzy logic. Three types of speech signal characteristics are used, and namely: energy, zero crossing rate and cepstral coefficients. Vectors of characteristics are normalized along the time axis by means of Dynamic Time Warping algorithm and supplied to the input of fuzzy logic system which produces the decision on an input speech signal belonging to one of vocabulary words.

Also the method of Fuzzy Vector Quantization realized in neural network base [4] can be used for isolated speech words recognition. Here the speech signal character is represented by the set of vectors of linear prediction coefficients.

The most perspective method applying fuzzy logic is based on a Fuzzy Pattern Matching method [5]. It states that the identifying feature of a speech signal is the history of structure changing with resonant signal frequencies. The system of isolated speech word recognition based on this method demonstrates the high level of recognition with the set of Japanese, German and English words [5]. In our opinion the head problem of Fuzzy Pattern Matching method is that of temporal alignment of comparing speech patterns. In the process of the problem solution there were proposals based on the application of linear temporal alignment [5] and nonlinear temporal alignment by means of Dynamic Time Warping algorithm [6]. Each of those algorithms has deficiencies affecting the quality of speech patterns' recognition. Thus the algorithm of linear alignment does not allow temporal non-uniformity of speech signal while the algorithm of nonlinear time alignment needs quite a lot of time for computation and reduces the distinguishability of speech patterns belonging to various classes. Therefore, the new algorithm named Optimal Temporal Alignment is proposed for time scale adjustment of comparing patterns. The alignment optimality criterion is maximization of similarity degree of comparing patterns [7].

2. Recognition system architecture

The general architecture of isolated words recognition system based on the Fuzzy Pattern Matching method is shown in figure 1. The recognition system input is the speech signal in the amplitude-time form while its output is the number of recognized words in vocabulary.

2.1. Deriving Informative Characteristics of a Speech Signal

A speech signal is converted to the two-dimensional spectral-time pattern (STP) which can calculated by means of signal distribution by frequencies in the group of 15 band-pass filters. Frequencies of the analysis are distributed in the interval of 200 … 5000 Hz with 1/3 octave step the Q-factor of each filter being equal to 6. Output signals of the filters are smoothed and sampled by 10 milliseconds. The pattern received reflects time dependent amplitudes alteration for prescribed frequency components of a speech signal and evaluates features of speech peculiarities perfectly well [5]. As we know a person pronounces words using a speech organ for resonant frequencies alteration. So, what is especially important information in STP is the structure of resonant frequencies i.e. local surges [5]. Consequently, it is possible to transform STP into a binary type preserving informative characteristics of speech by means of the following substitution: 1 – in the point of a local surge, 0 – in other points. The resulting

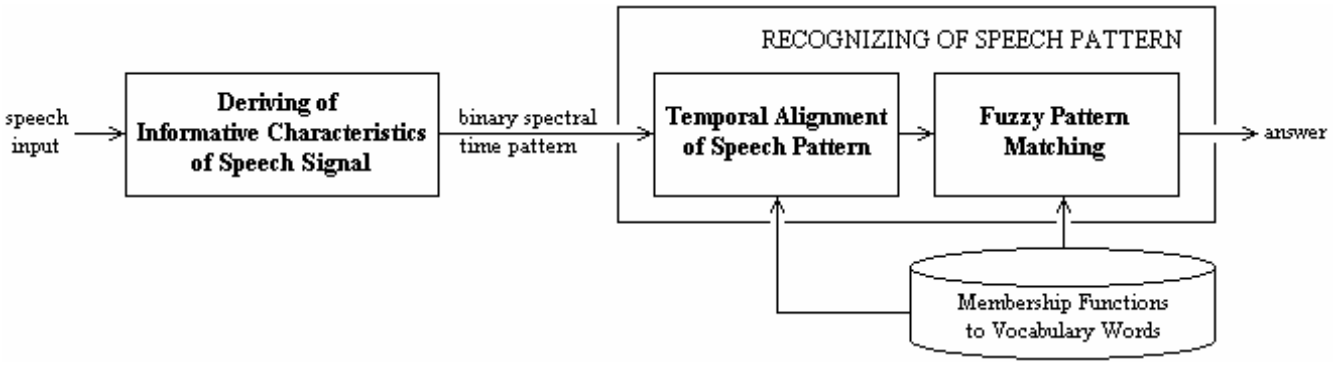

Figure 1: General architecture of isolated word recognition system

image is named a binary spectral-time pattern (BSTP) and used as features' reflection of a speech signal.

Let us consider the problem of local surges search in STP. $Xt(f)$ is the signal spectrum on frame t, $Yt(f)$ is the indication vector of local surges existence in the signal spectrum on t, f=1 … L frame. Let us determine numbers of $B = \{b_i \mid i \in \overline{1,M}\}$ local maximums and $S = \{s_j \mid j \in \overline{1,N}\}$ numbers of local minimums in $Xt(f)$ sequence. For each number of bi local maximum a xi value can be determined as the angle between two straight lines first of which is developed from a local maximum point $(b_i, X_t(b_i))$ to the left nearest local minimum point $(s_j, X_t(s_j))$ while the second is developed from the same local maximum point to the right nearest local minimum point $(s_{j+1}, X_t(s_{j+1}))$ (figure 2). We believe that those of determined local maximums bi values for which the rates of xi acuteness angles do not exceed the preset threshold α are local surges in the signal spectrum the xi angles being acuteness angles of these surges.

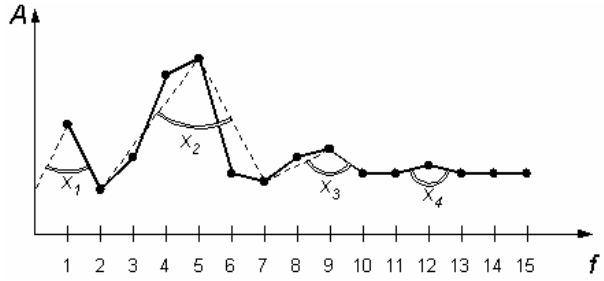

Figure 2: Example of the acuteness angles xi determination in the spectrum

Thus, the points of local surges in the signal spectrum $Yt(f)$ are determined as follows:

$$Y_t(f) = \begin{cases} 1, & f \in \{b_i \mid i \in \overline{1,M}\} \,\&\, x_i \le \alpha \\ 0, & \text{otherwise} \end{cases}$$

where $0 < \alpha \le 180^\circ$ is coefficient selected experimentally.

2.2. Optimal Temporal Alignment of Speech Pattern

Different speech patterns' duration belonging to the same class can differ considerably (for example, see figure 3). It can be explained by a speaker's speech tempo instability due to the effect of intonation, accent, etc. Therefore with matching patterns there arises the problem of temporal alignment of BSTP and a vocabulary pattern which are not equal in duration, i.e. their reduction up to the same length along the time axis. To provide the procedure execution a linear alignment algorithm for patterns reduction up to the equal length by means of uniform decimations or insertions was proposed in [5]. The shortcoming of this algorithm is that it ignores the non-uniform character of speech signal's course in time. Paper [6] proposes a nonlinear temporal normalization for comparing patterns' length alignment realized on the ground of dynamic programming. However such approach requires much computing time, besides, it reduces the distinguishability of speech patterns belonging to different classes. Therefore a new algorithm that is named Optimal Temporal Alignment is proposed [7] to extend comparing patterns up to identical length

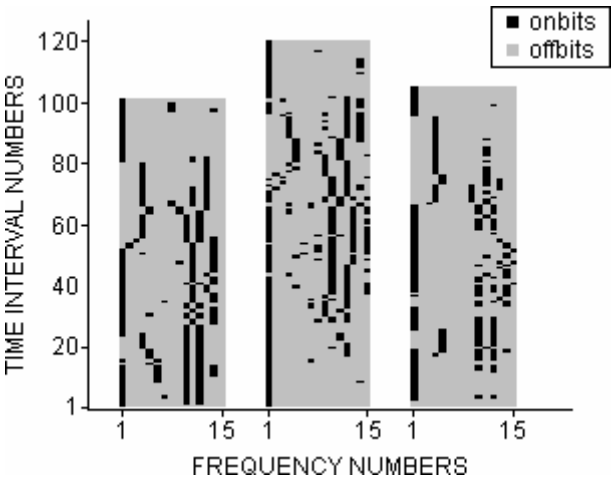

Figure 3: Binary spectral-time patterns of different records of the same Russian word [bolshe]

Let us deal with the problem of optimal temporal alignment of comparing patterns rj and y of different duration.

Let Tr be the size of a reference pattern (the fuzzy relation) $rj(f, t)$ along time axis, Ty be the size of an input pattern $y(f, t)$ along time axis, and $Tr \neq Ty$; F be the size of each pattern along frequency axis. $\tilde{r}_j(f,t,h)$ and $\tilde{y}(f,t,h)$ are patterns $rj(f, t)$ and $y(f, t)$ which are brought to the identical length in the following form:

$$\tilde{r}_j(f,t,h)=\begin{cases} r(f,t), & T_r > T_y \\ \hat{r}(f,t,h), & T_r < T_y \end{cases}$$

$$\tilde{y}(f,t,h)=\begin{cases} \hat{y}(f,t,h), & T_r > T_y \\ y(f,t), & T_r < T_y \end{cases}$$

where $\hat{r}(f,t,h)$ and $\hat{y}(f,t,h)$ are patterns expanded by the following function generation:

$$\hat{\varphi}(f,t,h)=\begin{cases} 0, & 1 \le t \le h \\ \varphi(f,t-h), & h < t \le T-(|T_r-T_y|-h) \\ 0, & T-(|T_r-T_y|-h) < t \le T \end{cases}$$

$$\hat{\varphi} \in \{\hat{r},\hat{y}\}\,;\varphi \in \{r,y\}$$

$$h \in [1,|T_r-T_y|]$$

The problem of optimal temporal alignment belongs to the class of nonlinear optimization problems of integer argument function and is represented as follows:

$$S_j(h)=\frac{\sum_{t=1}^{T}\sum_{f=1}^{F}\tilde{r}_j(f,t,h)\wedge\tilde{y}(f,t,h)}{\sum_{t=1}^{T}\sum_{f=1}^{F}\neg\tilde{r}_j(f,t,h)\wedge\tilde{y}(f,t,h)} \longrightarrow \max_h$$

$$1 \le h \le |T_r-T_y|$$

Optimal Temporal Alignment is used either to recognize the pair alignment of input pattern and each of vocabulary patterns or to train, i.e. membership functions development.

2.3. Fuzzy Pattern Matching

Recognition of isolated speech words is fulfilled on the ground of a Fuzzy Pattern Matching method [5]. BSTP of a speech word is the binary relation between *F* set (numbers of frequencies *f* on which the spectral analysis of a speech signal is made), and *T* set (numbers of time intervals *t*, on which speech signal is sampled by time frames) of the form:

$$f \in F, t \in T: \quad F \quad R \quad T$$

This binary relation specifies either presence or absence of local surge on *f* frequency in *f* time moment in a speech word. As soon as modifications in structure of the local surges caused by changes of intonation, speaking tempo, etc., are typical for different pronunciations of the same word then to describe a vocabulary speech pattern one would need a fuzzy relation *R* putting the value of membership function $\mu_R(x,y) \in [0,1]$ in conformity with each pair of elements $(f,t) \in F \times T$.

We denote the number of vocabulary words as *n*, the set of words as $I=\{i_1,i_2,...i_n\}$ and the set of fuzzy relations typical for each word as $R=\{r_1,r_2,...r_n\}$. Each fuzzy relation *rj* is formed as an arithmetical mean of reference BSTPs of *ij* word. The input unknown pattern *y* is considered to be a relation between the set of frequency numbers and that of time intervals. Then degrees of similarity *Sj* between the relation y and each fuzzy relation *rj* are calculated. The result of such recognition is the word *j* with response

$$j=\max_{j \in I}\{S_j\}$$

The degree of similarity is calculated using the following formula:

$$S_j=\frac{\int\int r_j(f,t)\wedge y(f,t)dfdt}{\int\int \neg r_j(f,t)\wedge y(f,t)dfdt}$$

3. Experimental Results

Here single-speaker and multi-speaker recognition experiments using the set of Russian words by the method of Fuzzy Pattern Matching with Optimal Temporal Alignment are described. The set included 105 commands of the text editor control, numerals from one to ten and ordinary words recorded by 8-bit samples in PCM format with 11,025 Hz frequency. 9 speakers, 6 men and 3 women, took part in the set composition each pronouncing all words thrice (see table 1).

Table 1. Speech database structure utilized in experiments.

	AGE		
GENDER	18 – 40	41 – 60	∑
male	5	1	**6**
female	2	1	**3**
∑	**7**	**2**	**9**
Each of 105 words was pronounced by each speaker 3 times. Total: 2835 records of the isolated words and word-combinations.			

Two of three deliveries of each word were used in single-speaker recognition for membership functions formation and one delivery for testing purposes. Experiments were conducted for each of nine speakers individually the results of recognition having been averaged for all speakers.

For multi-speaker recognition two of three deliveries of each word by each speaker were used for membership functions formation (18 speech deliveries total) while for testing one speech delivery by each speaker (9 speech deliveries total). To investigate the dependence between recognition quality and vocabulary size of recognizing system they also used the reduced sets of 52 and 20 words apart from the full set of 105 words.

In experiments 7 various values of α (threshold of acuteness angle of local surge in STP; this threshold is used for BSTP calculation – see Chapter 2) were used to determine an optimal α-value by a minimum criterion of recognition error.

Results of single-speaker recognition for various α values are presented in figure 4. It is found that the greatest portion of correctly recognized words 99.05 % corresponds to α=179.

Results of multi-speaker recognition for various α values and various vocabulary sizes of recognizing system are presented in figure 5. These results demonstrate that the best recognition quality 81.67 % is for 20 words vocabulary, 76.28% is for 52 words vocabulary, and 70.37 % is for 105

words vocabulary provided α value equals to 179. Besides, it is established, that the vocabulary size expansion is accompanied by the slight deterioration of multi-speaker recognition for each value of α.

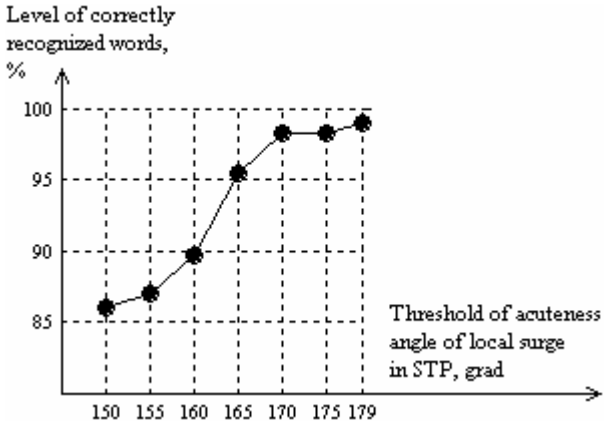

Figure 4: Results of single-speaker recognition for various threshold values of local surge acuteness angle in STP the vocabulary size being 105 words

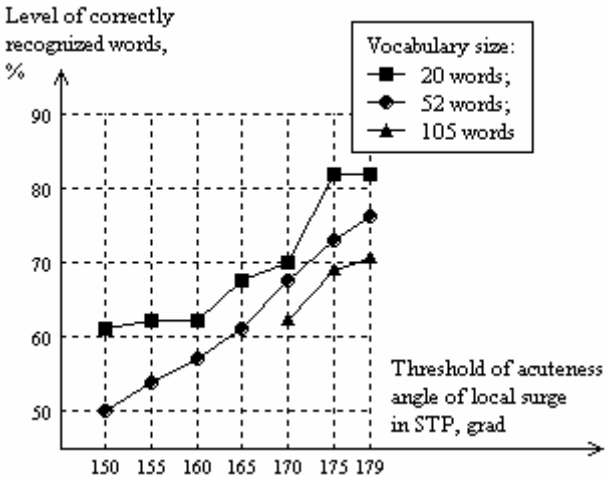

Figure 5: Results of multi-speaker recognition for various threshold values of local surge acuteness angle in STP

4. Conclusions

To recognize isolated words the Fuzzy Pattern Matching with Optimal Temporal Alignment is offered. The optimality criterion is similarity degree maximization of matched patterns. This algorithm application for single-speaker and multi-speaker recognition with the set of 105 Russian words including numerals and control commands by a text editor has shown good results in single-speaker recognition (99.05 % of correctly recognized words) and slightly lower results in multi-speaker recognition (70.37 % of correctly recognized words). When the vocabulary size of recognizing system was reduced to 20 words the multi-speaker recognition quality reached 81.67 %.

The results drawn from experiments are indicative of practical applicability of the offered isolated words recognition algorithm in the systems of speech commands control with preliminary adjustment to a speaker. Besides, on the ground of the offered algorithm one can create systems of speech command control without adjustment to a definite speaker though in such cases steady recognition requires smaller size vocabularies.

Further research will be directed to such speech characteristics set search and creation that would improve the invariance of Fuzzy Pattern Matching with Optimal Temporal Alignment to speaker voices' alteration. It would allow multi-speaker recognition systems development of great vocabulary size.

Besides, inclusion of the described algorithm of the isolated word recognition in the segmented-integrated model of speech perception is prospective. We believe, such approach will contribute in recognition reliability improvement using the mixture of experts where experts are parallel subsystems based on various principles of speech signal analysis [8, 9].

5. References

[1] Rabiner, L. and Juang, B. H., "Fundamentals of speech recognition", Prentice Hall, 1996.

[2] Zadeh, L. A., "Fuzzy Logic, Neural Networks and Soft Computing", Comm. of the ACM, vol. 37, no. 3, Mar. 1994.

[3] Beritelli, F., Cilia, G. and Cucè, A., "Small Vocabulary Word Recognition Based on Fuzzy Pattern Matching", Proc. of the European Symposium on Intelligent Techniques. Crete, Greece, 1999. Online: http://www.erudit.de/erudit/events/esit99/ 12651_p.pdf, accessed on 15 Apr 2008.

[4] HoseinNezhad, R., Moshiri., B and Eslambolchi, P., "Fusion of Spectrograph and LPC Analysis for Word Recognition: A New Fuzzy Approach", Proc. of the 7th International Conference on Information Fusion, pp. 449-454. Stockholm, Sweden, 2004. Online: http://www.fusion2004.foi.se/papers/IF04-0449.pdf, accessed on 15 April 2008.

[5] Asai,K.,Vatada,D.,Ivai,S. and others, "Speech recognition", in "Applied Fuzzy Systems", pp. 157-170. Translated into Russian, Moscow, Mir, 1993.

[6] Bondarenko, I.Yu. and Fedyaev, O.I., "The analysis of efficiency of the fuzzy pattern matching method for isolated word recognition", Proc. of the 6th International Conference "Intellectual Analysis of the Information" IAI 2006, pp. 20-27. Kiev, 2006. [in Russian]

[7] Fedyaev, O.I. and Bondarenko, I.Yu., "Fuzzy Pattern Matching with Optimal Temporal Alignment for Single-speaker and Multi-speaker Isolated Word Recognition", Informatics, Cybernetics and Computer Science, 8 (120): pp. 273-281. Donetsk National Technical University. Donetsk, 2007. [in Russian]

[8] Bondarenko, I.Yu, Gladunov, S.A., Fedyaev, O.I., "Segmented-integrated structure of the channel of speech control of program systems", Proc. of the 10th National Conference on Artificial Intelligence, pp. 841-849. Moscow, 2006. [in Russian].

[9] Bondarenko, I.Yu. and Fedyaev, O.I., "Segmentally-integral model for spoken pattern recognition based on functional brain asymmetry conceptions", Proc. of the 9th All-Ukrainian International Conference on Signal/Image Processing and Pattern Recognition UkrObraz'2008, pp. 81-84. Kiev, 2008. [in Ukrainian]