The New Algorithm for Speech Control in the Cockpit

Aydar Gabdrakhmanov[1,*]

[1] Moscow Institute of Physics and Technology (State University), Moscow, Russia

Abstract. Speech technologies are being developed intensively in the recent years, especially the automatic speech recognition as an additional input method in human interface and technical devices. Most of the known algorithms for speech control have small probability of correct recognition. Widespread methods, like Markov models and neural networks, which require large processing power, allow recognizing the words with a probability of no more than 85–92 %. Such accuracy is not enough to use the voice control on board of a modern aircraft. The article is devoted to a problem of improving the automatic speech recognition's accuracy. A version of word recognition algorithm based on the classical approach is suggested, it includes the comparison with the patterns. In this work to improve the recognition's accuracy a new method of calculating a similarity measurement between the recognizable word and the pattern, which based on z-Fisher transformation, is described. This article also contains an algorithm's modification that takes into account the fixed ratios with the patterns of other words and uses the words adjustment to the pattern with dynamic programming elements. The usage of fixed relations between words provides additional information, which positively affects the recognition. The experimental results of the developed algorithm's approbation on a large amount of speech data are presented.

1 Introduction

In recent years a lot of researches on the improvement of the cockpit interface based on modern audio technologies appeared, for example: control of onboard equipment based on automatic speech recognition [1], the creation of surround sound effect to enhance the information content of a sound indication [2], the analysis in order to ensure the safety characteristics of speech signals [3]. The creation of a cockpit voice interface is complicated by many factors: quality of the existing methods of recognition [1, 4, 5], where the percentage of correctly recognized words rarely reaches 90 - 95%; overload impact [6, 7]; presence of strong acoustic noise [8, 9]; occupational diseases of hearing organs of helicopter aviation crew[10].

2 Formulation of the problem

We propose some algorithms, which can increase the probability of correct recognition, which is critical when using voice control on board the aircraft, where any mistake affects the safety of the flight.

Our algorithms are based on the traditional recognition method [5], which contains the comparison the parametric portraits of words with patterns. Description of the algorithmic implementation given in the article [11]. The conceptual simplicity of this method is useful for evaluating the effectiveness of the proposed new algorithmic solutions. In order to increase the level of recognition we propose the following new approaches:

- A modified measure of closeness between parametric portrait of a recognizable word and patterns. The new measure is to find the middle of the z-Fisher transformation of the correlation coefficients between word's and patter's frames.
- Use of the information about fixed ratios of recognizable words, not only with the appropriate pattern, but also with all other members in the dictionary, which increases the probability of successful recognition.

The initial version of algorithm to the simple case of operator dependent recognition is shown in [12]. This paper proposes a final version of the algorithm and the experimental results in case of operator independent recognition. The effectiveness of the considered algorithms significantly increases when using adaptation of words in the time domain [13], based on dynamic programming approach [4, 14]. Algorithmic implementation is described in detail in [15].

3 Recognition based on comparison with standards

The algorithm is applied classical scheme of finding the maximum value of the measure of closeness between the

[*] Corresponding author: besha5500@mail.ru

recognizable word and all standards of the dictionary. Let us consider the version used in this study.

Let there be recording of the words in the time domain. For a single word such record has the form

$$x(t_i), i=1,2,\ldots,N_k, \qquad (1)$$

$x(t_i)$ – the amplitude of the microphone signal

N_k – number of discrete values of the speech signal.

In our case, the sampling frequency is f=20050 Hz, which corresponds to the sampling interval τ=1/f=1/22050~ 0,05 ms.

Apply to form (1) the spectral-time transformation, which divide the record in time scale into N_t intervals with duration 20…40 ms, on each of which, by using fast Fourier transform algorithm, windows Hannah and averaging over frequencies, compute $N_f = 30..40$ logarithms of signal density values [5, 11]. As a result, we get a parametric portrait of a word in the form of a matrix

$$\mathbf{X} = \{x_{ij}\}, i = 1,2,\ldots,N_f, j = 1,2,\ldots,N_t, \qquad (2)$$

where columns correspond N_t quantization intervals in time, and the rows contain values N_f frequency components. Let us apply to the matrix the notation $\{x_{ij}\}$.

In our experiments we use a uniform frequency scale and Mel scale, widespread in the analysis of speech signals [4, 5]:

$$m = 1127\ln(1+\frac{f}{700}),$$

here f – sound frequency in Hz, m – sound frequency in Mel scale.

To create the standard of the word take E implementations of this word and calculate the expected value of E parametric portraits

$$\mathbf{X}_E = \{x_{Eij}\} = \frac{1}{E}\sum_{k=1}^{E}\{x_{ij}^k\}, i = 1,2,\ldots,N_f, j = 1,2,\ldots,N_t \ (3)$$

The resulting portrait (3) we take as a standard of the word. If it is required to recognize M words, then using formula (3) form M corresponding standards

$$\{x_{Eij}^l\},\ l = 1,2,\ldots,M, \qquad (4)$$

where each standard is a matrix of dimension $N_f \times N_t$.

Let it be required to recognize a word that is received in time record (1). Then create the parametric portrait (2) and calculate the measure of proximity (distance) between portrait (2) and each of the standards (4). As the distance we choose, for example, estimate of the correlation coefficient:

$$\hat{r}_{lx} = \frac{\sum_{i=1}^{N_f}\sum_{j=1}^{N_t}\left(x_{ij}-\hat{m}_x\right)\left(x_{ij}^l-\hat{m}_l\right)}{\sqrt{\sum_{i=1}^{N_f}\sum_{j=1}^{N_t}\left(x_{ij}-\hat{m}_x\right)^2 \ \sum_{i=1}^{N_f}\sum_{j=1}^{N_t}\left(x_{ij}^l-\hat{m}_l\right)^2}}, \ (5)$$

$$\hat{m}_x = \frac{1}{N_f N_t}\sum_{i=1}^{N_f}\sum_{j=1}^{N_t} x_{ij};\ \hat{m}_l = \frac{1}{N_f N_t}\sum_{i=1}^{N_f}\sum_{j=1}^{N_t} x_{ij}^l, \ (6)$$

$$l = 1,2,\ldots,M,$$

$\hat{r}_{lx}$ – estimate of the correlation coefficient between recognizable word and standard number l,

$\hat{m}_x$ $\hat{m}_l$ – estimate of the expected value or mean element of the matrix of the portrait of recognizable word, portrait of standard number l,

x_{ij}^l – the element of portrait of standard number *l*, located at the intersection of row number *i* and column number *j*.

The result of recognition is defined by the maximum correlation coefficients among all M standards

$$\hat{r}_{l_{\max}x} = \max_l \hat{r}_{lx}, , \qquad (7)$$

$$l = 1,2,\ldots,M$$

$l_{\max}$ – index corresponding to the maximum correlation coefficient and determining the recognition result.

There are other distance measures such as Euclidean distance between the matrices of parametric portraits, but experience shows that they give similar results [1]. Therefore, we will consider this algorithm as a base.

4 Recognition on the basis of comparison with other standards

Considered the traditional recognition scheme is widespread and is to find a standard with a maximum measure of closeness to recognizable word. It reflects the natural process of understanding speech in which a person chooses the most appropriate standard to recognizable word, and does not care about the correlations with other words. This scheme is based on the assumption that a measure of closeness to the word with "its" standard more than with others or "foreign" standards in the dictionary.

When machining in order to increase the probability of correct recognition we offer other schemes of comparison. It is advisable to use the additional information contained in fixed ratios between different words. Note that such a change in the formulation of the problem should not lead to a significant increase in the amount of calculation, as discussed above the detection scheme still requires a comparison with all the standards. Let formulate the new recognition algorithm that uses the results of comparison with all the standards, which are discarded after a maximum location in a traditional scheme.

As part of the basic algorithm (1) - (7) it is advisable to additionally take into account the values of the correlation coefficients between the different words.
Let the training database contains M words with E implementation of each. We form M standards in accordance with (4).

Further, according to (5) for each k word, $k = 1, 2, \ldots, M$, compute the average by E implementations correlation coefficients with all standards:

$$\bar{r}_{lk} = \frac{1}{E}\sum_{i=1}^{E}\hat{r}_{lk}(i)\,, \qquad (8)$$

$$k = 1, 2, \ldots, M,\ \ l = 1,\, 2,\, \ldots, M\,,$$

k , l – codes corresponding to the words and standards; $\hat{r}_{lk}(i)$ – estimate of the correlation coefficient between l standard and i implementation of k word; E – the number of realizations of each word in the training database.

Equation (8) defines the matrix of the average coefficients **R** with the size $M \times M$

$$\mathbf{R} = \{\bar{r}_{lk}\},$$

$$k = 1,\, 2,\, \ldots, M,\quad l = 1,\, 2,\, \ldots, M$$

in which, for example, the second column consists of average correlation coefficients with all standards and the second word. It turns out that each k word is associated with k column of the matrix $\mathbf{R}$, or vector with dimension M of average correlation coefficients $\mathbf{R}_k = \{\bar{r}_{lk}\}, l = 1, 2, \ldots, M$.

In recognition of unknown words x calculate from the formula (5) evaluation of the correlation coefficients of its parametric portrait (2) with all M standards, so obtain the vector with dimension M :

$$R_x = \{\hat{r}_{lx}\},\ . \qquad (9)$$

$$l = 1, 2, \ldots, M$$

The result of recognition is defined by the minimum norm of subtraction $d_{k_{\min}}$ between the vector (9) and the columns of the matrix $\mathbf{R}$

$$d_{k_{\min}} = \min_k \left\| \{\hat{r}_{lx}\} - \{\bar{r}_{lk}\} \right\| . \qquad (10)$$

Taking the Euclidean norm in the formula (10), we obtain

$$d_{k_{\min}} = \min_k \left\{ \sum_{l=1}^{M} \left(\hat{r}_{lx} - \bar{r}_{lk} \right)^2 \right\} . \qquad (11)$$

In equations (10) and (11) we search the minimum by all words in the dictionary $k = 1, 2, \ldots, M$.

5 Modified proximity measure

We create the modified measure of closeness between the recognizable words and standards, which will significantly increase the probability of correct recognition.

Estimates of correlation coefficients have a special asymmetrical distribution [14], so in statistical relation more stable results can be obtained by substituting in formulas (8) - (11) z-Fisher transformation instead of the correlation coefficients, having approximately normal distribution

$$z = \frac{1}{2}\log\left(\frac{1+\hat{r}}{1-\hat{r}}\right), \qquad (12)$$

$\hat{r}$ – evaluation of the correlation coefficient.

Furthermore, let calculate the correlation coefficient not around the hall parametric portrait of the word and its standard as the base in the formula (5), but separately for each time interval (frame) followed by averaging values. Then, for partition to the number of frames N_t calculate for each j frame of recognizable word an assessment of the correlation coefficient with *j* frame of *l* standard, going through all the standards $l = 1, 2, \ldots, M$. In this case, instead of the expression (5) apply the following formula

$$\hat{r}_{lxj} = \frac{\sum_{i=1}^{N_f}\left(x_{ij} - \hat{m}_x\right)\left(x_{ij}^{l} - \hat{m}_l\right)}{\sqrt{\sum_{i=1}^{N_f}\left(x_{ij} - \hat{m}_x\right)^2 \sum_{i=1}^{N_f}\left(x_{ij}^{l} - \hat{m}_l\right)^2}}\,, \qquad (13)$$

$$\hat{m}_x = \frac{1}{N_f}\sum_{i=1}^{N_f} x_{ij}; \qquad \hat{m}_l = \frac{1}{N_f}\sum_{i=1}^{N_f} x_{ij}^{l}\,,$$

$$j = 1, 2, \ldots, N_t\,,\ \ l = 1, 2, \ldots, .$$

$\hat{r}_{lxj}$ –- evaluation of the correlation coefficient between *j* frame of recognizable word with *j* frame of *l* standard.

Then go to the Fisher z-transform. For this purpose, in the formula (12) substitute estimate of the correlation coefficient (13):

$$z_{lxj} = \frac{1}{2}\log\left(\frac{1+\hat{r}_{lxj}}{1-\hat{r}_{lxj}}\right), \qquad (14)$$

$$j = 1, 2, \ldots, N_t\,, l = 1, 2, \ldots, M\ .$$

As a final measure of closeness take the average over all intervals N_t z-Fisher transformation (14) of the correlation coefficients (13) between frames of word and the reference:

$$\bar{z}_{lx} = \frac{1}{N_t}\sum_{j=1}^{N_t}\hat{z}_{lxj}\,, \qquad (15)$$

$$i = 1, 2, \ldots, N_f\,, l = 1, 2, \ldots, M\ .$$

The proposed two new algorithms 8-11 and 12-15 is applicable in conjunction with the traditional method of pre-adjusting recognizable words in length scale, which reduces the effect of variations in pronunciation of different speakers [10]. The method uses dynamic programming [4, 14, 15].

6 Experimental evaluation of the effectiveness of the developed algorithms

Recognition results of proposed algorithms are tested on the record material of seven different speakers and dictionary with 20 isolated words. Every speaker pronounced 600 words (20 words and 30 realizations of each). Standards formed using similar records of the eighth speaker, so we discuss here speaker independent case of recognition with small learning base composed only one speaker. Recognition results for all speakers and each algorithm are shown in table 1. Where 1 - a basic recognition with a measure of the closeness - the correlation coefficient (6); 2 - a modified proximity measure with averaged z-Fisher transformation (15) of the correlation coefficient; 3 - similar to option 2, but with the implementation of pre-adjusting words in length scale, 4 – the comparison with the standards of "foreign" words, and using modified proximity measure, and pre-adjusting words in length scale.

Table 1. Recognition results (the number of errors in %)

Algorithm	**Speaker**							**The average error rate**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
1	8,2	15,0	12,2	15,0	11,8	8,8	8,0	**11,3**
2	667	9,8	5,5	6,5	9,5	6,0	4,2	**6,9**
3	6,8	7,2	1,0	4,7	5,7	6,0	1,2	**4,6**
4	5,7	5,3	2,0	2,2	4,0	3,5	1,7	**3,5**

Note that each of the listed above algorithmic approach reduces the average error rate. For example, a modified proximity measure reduces the number of errors by 4.4%, pre-adjusting words in length scale together with a modified measure reduce the failure rate by 2.3%, comparing with the "foreign" standards improves results further by 1.1%. Thus, the combined use of three algorithms (modified proximity measure, duration adjustment, comparison with "foreign" standards) allow us to recognize correctly 96.5% of all words. Note that in this experiment we used a small training base composed only one speaker.

7 Conclusions

New approaches to improve the quality of voice recognition are developed:
- a modified measure of closeness between parametric portrait of a recognizable word and patterns; the new measure is to find the middle of the z-Fisher transformation of the correlation coefficients between word's and patter's frames;
- recognition based on a comparison with the "foreign" standards, using fixed relations between different words.

Experimental results show the effectiveness of the proposed algorithms, especially when combined with the method for adjusting the word's length based on dynamic programming.

To the dictionary with 20 isolated words and 30 realizations for each and for seven speakers in a speaker-independent version the combined use of the considered algorithms shows the result up to 96.5% correct recognitions. In this study, the possibility of increasing the quality of recognition, based on the expansion of the training base, deliberately not been used for a better estimation of the effectiveness of the proposed algorithms.

References

1. V. Soshnikov et al., *Trudy GosNIIAS. Seriia: Voprosy avioniki*, **1**, 24–32 (2016)
2. G. Sebryakov et al., *Modelirovanie aviacionnyh sistem: sb. Dokl*, **3**, 452–458 (2011)
3. G. Bondaros et al., *Vestnik komp'iuternykh i informatsionnykh tekhnologii*, **11**, 2–11 (2009)
4. A. Peinado, J. Segura *Speech recognition over digital channels: Robustness and Standards* (New Jersey, John Wiley&Sons, 2006)
5. L. Rabiner, B. Juang, Fundamentals of speech recognition englewood Cliffs (New Jersey, Prentice-Hall International, 1993)
6. A. Ivanov et al., *Vestnik komp'iuternykh i informatsionnykh tekhnologii*, **5**, 3–7 (2012)
7. A. Ivanov et al. *Vestnik komp'yuternyh i informacionnyh tekhnologij*, **6**, 3–7 (2012)
8. V. Chuchupal et al., *Nauka i obrazovanie: nauchnoe izdanie MGTU im. N. Baumana*, **1**, 103–114 (2013)
9. A. Gabdrakhmanov et al., *Nauchnye chteniya po aviacii, posvyashchennye pamyati N. E. ZHukovskogo*, **1**, 330–334 (2013)
10. A. Ivanov et al., *Medicina truda i promyshlennaya ehkologiya*, **11**, 40–45 (2014)
11. A. Gabdrakhmanov et al., *Mekhatronika, avtomatizatsiia, upravlenie*, **9**, 599–604 (2015)
12. O.N. Korsun, A. Gabdrakhmanov, *Nauchnye chteniya po aviacii, posvyashchennye pamyati N. E. ZHukovskogo*, **4**, 331–336 (2016)
13. O.N. Korsun, A. Gabdrakhmanov, *Nauchnye chteniya po aviacii, posvyashchennye pamyati N. E. ZHukovskogo*, **3**, 184–187 (2015)
14. E.S. Wentzel, *Operations research* (DROFA, 2004)
15. O.N. Korsun, A.V. Poliev, *Journal of Computer and Systems Sciences International*, **55**(4), 115–124 (2016)