Optimal pattern synthesis for speech recognition based on principal component analysis

O N Korsun[1 *] and A V Poliyev[2]

[1]State Research Institute of Aviation Systems, Moscow, 125319, Russia
[2]Moscow Institute of Physics and Technology, Moscow, 117303, Russia

E-mail: marmottoduo@yandex.ru

Abstract. The algorithm for building an optimal pattern for the purpose of automatic speech recognition, which increases the probability of correct recognition, is developed and presented in this work. The optimal pattern forming is based on the decomposition of an initial pattern to principal components, which enables to reduce the dimension of multi-parameter optimization problem. At the next step the training samples are introduced and the optimal estimates for principal components decomposition coefficients are obtained by a numeric parameter optimization algorithm. Finally, we consider the experiment results that show the improvement in speech recognition introduced by the proposed optimization algorithm.

1. Introduction

A reliable and rational man-machine interface is an important task of modern engineering [1-3]. In recent years in avionics we witness a considerable growth of attention to audio interfaces and speech recognition [4-10], since their successful implementation may improve the flight safety and reduce the pilots' workload [1,10]. For example, on Eurofighter Typhoon since 2005 a speaker-dependent system based on the comparing with the patterns is employed.

The principal requirement for speech interface in avionics is high probability of correct word recognition even under strong noises, because an error in aircraft control may be critical for the flight safety. In addition, the time complexity of the algorithm is an important point for onboard software.

For the recognition of speech commands, the method of comparison with a pattern shows a good performance for speech control of aviation equipment. In this work, one proposes the new method for a pattern generation and its improvements using the principal component method. The results of word recognition with new patterns are presented in the experimental part. And, in conclusion, one considers a way to simplify the algorithm in order to reduce the working time while preserving the results.

2. Speech signal parametric portrait in automatic speech recognition

The traditional methods for automatic recognition of speech commands are based on the time-spectral quantisation of an input word record. The transform used in this article obtains a time-spectral parametric portrait and includes the following steps. In the beginning, a speech signal is divided into equal intervals of 10-30 ms. After that, one uses the standard digital signal processing procedures, such as amplification of high-frequency signal components, interval weighing by Hannah window, fast Fourier transform, averaging over frequencies and logarithm of spectral densities [11]. Each of these intervals is projected into 30-40 frequency bands [1]. As a result, for each of the time intervals bands we obtain the estimates of spectral densities logarithms in 30-40 frequency bands [4,8,10]

These parameters, presented in a form of a matrix, we call the parametric portrait of the word. The columns of this portrait characterize the spectral decomposition of the speech signal for each time interval.

Further, the pattern parametric portrait is compiled from available parametric portraits of the speech commands records. The simplest way to obtain a pattern is to use the mean of the parametric portraits of several similar records. Then, the parametric portrait of the recognized word is compared with the pattern parametric portrait. This comparison may be carried out by the criterion of the maximum correlation between the elements of parametric portrait matrices. After comparing the recognized word with all patterns, the pattern that provides the correlation maximum is selected. The speech command associated to this pattern is recognized as the result of automatic recognition.

3. Principal component analysis

The principal component analysis (PCA) is traditionally used to reduce the number of dimension of the observed vectors space without significant loss of information. The PCA method applications cover a great number of fields from statistics to medicine [12]. The principal components are the orthogonal vectors that form the orthogonal coordinates system in a space of observed data.

Let us consider an initial set of vectors X in a linear space $\boldsymbol{L}^{p}$. The principal component analysis enables to switch to a basis $\boldsymbol{L}^{p'}, (p' \ll p)$ such that the first component (the first vector of a basis) corresponds to the direction along which the variance of the initial set of vectors is maximal. The direction of the second component (the second vector of a basis) is selected so that the variance of the original vectors along it is maximal provided that it is orthogonal to the first vector of a basis. The remaining basis vectors are defined similarly. As a result, the directions of the basis vectors, that are selected to maximize variance of the initial set along the first components, are called the principal components. It turns out that the main variability of the initial set of vectors is presented in the first few components, and it is possible to move to the space of smaller dimension discarding the remaining (less significant) components [12].

Let us consider a multi-dimensional random observation $X = \begin{pmatrix} x^{(1)} \\ \vdots \\ x^{(p)} \end{pmatrix}$. The problem is to reduce the dimension of X from p to p' with no considerable loss of information. One can solve this problem by defining all possible linear combinations of orthogonal normalized combinations of initial observations $z^{(j)}(X) = c_{j1}(x^{(1)} - \mu^{(1)}) + \cdots + c_{jp}(x^{(p)} - \mu^{(p)})$, where $[\mu^{(1)}, \dots, \mu^{(p)}]'$ - vector of observation X means.

A measure of information value for p'-dimensional system of $(z^{(1)}(X), \dots, z^{(p')}(X))$ may be defined as $I_{p'} = \frac{Dz^1 + \cdots + Dz^{p'}}{Dx^1 + \cdots + Dx^{p}}$, where Dz is the operation calculating the random variable variance. It can be shown that all p principal components of X may be represented as $Z = LX$, where $Z = (z^{(1)}, \dots, z^{(p)})', X = (x^{(1)}, \dots, x^{(p)})'$, and matrix L consists of rows $l_j = (l_{j1}, \dots, l_{jp}), j = \overline{1, p}$, which are the eigenvectors of the random variable X covariance matrix, $\sum = \sigma_{ij}, i, j = \overline{1, p}$. The components are enumerated according to decrease of corresponding eigenvalues $\lambda_j, j = \overline{1, p}$.

The basic properties of principal components:

- L is an orthogonal matrix, that is $LL' = L'L = I$, where I – identity matrix;
- The covariance matrix of the principal components vectors $\sum_Z = L\sum L' = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix}$;
- The sum of observation components $x^{(1)}, \dots, x^{(p)}$ variances is equal to the sum of the all principal components $z^{(1)}, \dots, z^{(p)}$ variances;
- Criterion of PCA information value is $I_{p'}(Z(X)) = \frac{Dz^1 + \cdots + Dz^{p'}}{Dx^1 + \cdots + Dx^{p}} = \frac{\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2 + \cdots + \boldsymbol{\lambda}_{p'}}{\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2 + \cdots + \boldsymbol{\lambda}_{p}}$, where $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \dots, \boldsymbol{\lambda}_p$ – the eigenvalues of the covariance matrix $\sum$ of the vector X arranged in decreasing order. One can use this criterion as a reference point of view to make decide about the principal components to be withdrawn with no considerable damage to the information content, thereby reducing the dimension of the space used.

4. Development and validation of the algorithms for spectral decomposition of a word portrait on principal components

Let us consider parametric portraits of M different samples of the same word $\{x_{lj}(k)\}, k = 1, 2, \dots, M; l = 1, 2, \dots N_t; j = 1, 2, \dots, N_f$. Let us transform a matrix portrait for each k into a one-dimensional array with a

number of elements equal to $P = N_f N_t$. Therefore, we have M vectors each with P dimensions $\{x_{ik}\}, k = 1, 2, \dots, M; i = 1, 2, \dots, P$. Explicitly:

$$x_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \dots \\ x_{p1} \end{bmatrix}, x_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \dots \\ x_{p2} \end{bmatrix}, \dots, x_M = \begin{bmatrix} x_{1M} \\ x_{2M} \\ \dots \\ x_{pM} \end{bmatrix} \quad (1)$$

Let us assemble these M vectors into a matrix of dimension $p \times M$:

$$X = [x_1 x_2 \dots x_M] = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \dots & \dots & \dots & \dots \\ x_{p1} & x_{p2} & \dots & x_{pM} \end{bmatrix} \quad (2)$$

Further we may estimate the matrix of correlation moments for vectors $x_1, x_2, \dots, x_M$ of dimension M

$$K_x = X^T X = \begin{bmatrix} x_1^T x_1 & x_1^T x_2 & \dots & x_1^T x_M \\ x_2^T x_1 & x_1^T x_2 & \dots & x_2^T x_M \\ \dots & \dots & \dots & \dots \\ x_M^T x_1 & x_M^T x_2 & \dots & x_M^T x_M \end{bmatrix} \quad (3)$$

One may note that each element of the matrix K_x is a scalar product of the corresponding vectors and so matrix K_x is symmetrical. Therefore, it is possible to calculate M eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_M$ and corresponding M-dimensional eigenvectors $l_1, l_2, \dots, l_M$ for this matrix. The eigenvalues may be arranged in the descending order $\lambda_1 > \lambda_2 > \dots > \lambda_M$.

The first principal component z_1 is defined as a linear combination of the original vectors $x_1, x_2, \dots, x_M$ with coefficients equal to the elements of the first eigenvector:

$$l_1^T = [l_{11} l_{21} \dots l_{M1}]: z_1 = l_{11} x_1 + l_{21} x_2 + \dots + l_{M1} x_M = \sum_{i=1}^{M} l_{i1} x_i \quad (4)$$

Similarly, the principal components $j = 2, 3, \dots, M$ are calculated using the formula:

$$z_j = \sum_{i=1}^{M} l_{ij} x_i \quad (5)$$

According to the theory described above the total principal components variance is equal to the total variance of the initial vectors, i.e. transform does not alter the energy of the signal. The variance of the principal components z_j is equal to the corresponding eigenvalue λ_j. So the behavior of the system is determined mainly by the first few principal components z_j, $j = 1, 2, \dots, M'$. It allows us to reduce the dimension of the problem and consider M' principal components instead of the M initial vectors.

To estimate the errors introduced during the transition from the original vectors M to M' principal components one can use the expression:

$$I_p = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_{M'}}{\lambda_1 + \lambda_2 + \dots + \lambda_M}. \quad (6)$$

It is clear that the information value criterion (6) is equal to the ratio of variance, explained by the M' selected components, to the total variance of the initial signal.

5. Development of the algorithm for optimal patterns synthesis based on principal components analysis

In this work, the recognition is performed for three Russian words: "pilotazh", "masshtab" and "navigatsiya", which are pronounced by ten different speakers without noise and by four speakers with noise in the headphones.

The noise in the headphones is implemented as follows. The speaker puts the headphones on, which play the noise of the given volume. Also the speaker can hardly hear his own voice, i.e. there is no feedback in the headphones. The noise is not recorded along with the spoken word, and is only played in the headphones. This procedure is performed not to analyze the influence of noise on the recognition process, but to evaluate the effect of noise on changes of the speaker's speech.

The normal patterns for the recognition of each of the three words were formed as mean parametric portraits of ten speakers with no noise. It is known that the noise in the headphones changes the speech, so the normal patterns formed in no noise conditions are not fully adequate to the records of speakers with noise

in the headphones. The idea of the experiment was to adjust normal patterns to the data with noise in the headphones using principal components, numeric optimization and additional training samples.

The principal components were formed as follows. Speech samples of ten speakers with no noise were introduced to build ten separate pattern, one pattern vector for each speaker. The parametric portraits of these $M = 10$ patterns were considered further as original vectors $x_1, x_2, \dots, x_M$ (1). Then according to (2)-(5) $M = 10$ principal components were built: z_j , $j = 1,2, \dots, M$. Then the formula (6) was applied which enabled to select $M' = 6$ components at information value criterion (6) equal to 0.95. At the next step, we expressed the normal patterns as a linear combination of selected principal components:

$$e_{pat\_norm} = k_0 Z_0 + k_1 Z_1 + \dots + k_6 Z_6 . \quad (7)$$

The coefficients $k_0, \dots, k_6$ were obtained through the multiple linear regression; the coefficient k_0 was added to incorporate the mean of normal pattern. Naturally, the decomposition (7) was applied separately to each of normal patterns that is the patterns for three words in consideration. Now we should note that usually a parametric portrait of a word (recognition pattern is also a parametric portrait) is a vector that includes 1200-1800 elements. So the PCA enabled us to reduce the set of parameters to a few coefficients $k_0, \dots, k_6$.

For normal pattern optimization it is necessary to form an optimization criterion and a training sample. The natural criterion for a search of coefficients $k_0, \dots, k_6$ is the minimum of recognition errors: $N_{errors} \to min$. The fault is that the criterion values are discreet. The problem of three words recognition is quite simple, so the number of errors at training sample may be inconsiderable or even equal to zero. As a result, the discrete criterion may not be sensitive enough. In this research we used the continuous scalar measure of recognition quality ΔF. This measure was introduced in [13]. For each recognized word the measure ΔF is equal to Fisher z-transforms difference between maximum correlation and correlation nearest to maximum. That is, the greater is distinction between the «winner» word and «the best of the rest» words, the higher is the recognition quality.

For training sample we selected one out of four speaker records with noise in the headphones. More precisely, we chose N-v speaker records with the 80 dB noise in the headphones. The search of coefficients $k_0, \dots, k_6$ was performed by a widely known coordinate descent method. To achieve an extreme point 10-15 iterations proved to be enough. The optimization algorithm was applied to the patterns of all the three words, so the optimization problem dimension was equal to 21. For the starting values of the optimized parameters we assumed the coefficients of normal patterns linear decomposition (7).

6. Results

To estimate the optimized pattern efficiency we performed the automatic recognition tests. At this step we used all four speakers records with two levels of noise in the headphones: 80 dB and 90 dB, so the number of testing sets was equal to eight. The results of the optimized pattern application have shown the improvement in samples recognition not only for the N-v words, but also for other speakers outside the training set. The total number of errors before optimization for the three words was equal to 3.15% and after optimization has reduced to 0.89%. The results are shown in Tables 1 and 2.

Table 1. The samples recognition results with the 80 dB noise in the headphones with normal (1) and optimized (2) patterns.

Speaker	Word	Errors 1	Errors 2
B-k	pilotazh	6	0
	masshtab	0	0
	navigatsiya	0	8
G-v	pilotazh	3	0
	masshtab	0	0
	navigatsiya	0	0
N-v	pilotazh	8	0
	masshtab	0	0
	navigatsiya	0	0
F-v	pilotazh	9	0
	masshtab	0	0
	navigatsiya	0	0

Table 2. The samples recognition results with the 90 dB noise in the headphones with normal (1) and optimized (2) patterns.

Speaker	Word	Errors 1	Errors 2
B-k	pilotazh	7	0
	masshtab	0	0
	navigatsiya	0	1
G-v	pilotazh	3	0
	masshtab	0	0
	navigatsiya	0	2
N-v	pilotazh	12	0
	masshtab	0	0
	navigatsiya	0	0
F-v	pilotazh	4	0
	masshtab	0	0
	navigatsiya	8	4

7. Conclusion

The patterns, obtained by optimizing the principal components coefficients, showed significantly fewer errors in the recognition for the majority of samples. This method proved to be the an effective way to improve word recognition results, along with the words splitting into homogeneous parts [10], which also gives improved results for this problem. The achieved positive results confirm validity of proposed the method, which includes PCA based reduction of parameter space dimension and pattern optimization. The further progress is to be obtained by solving recognition problems with greater vocabulary using more sophisticated optimization algorithms [14]

Acknoledgements

The research is supported by Russian fund for basic research (RFBR), project 15-08-06946-a.

References

[1] Evdokimenkov V.N., Kim R.V., Krasil'shchikov M.N., Sebryakov G.G. 2015 *Journal of Computer and Systems Sciences International*. 54(4) 609
[2] Polyak B T and Khlebnikov M V 2017 *Automation and Remote Control* 3 130
[3] Kolokolov A.S. 2006 *Problemy upravleniya* 3 13–18 (In Russian)
[4] Rabiner L.R., Juang B.-H. 1993 *Fundamentals of Speech Recognition* (Englewood Cliffs. NJ: Prentice-Hall)
[5] Springer Handbook on Speech Processing and Speech Communication 2007 (Berlin, Heidelberg: Springer)
[6] Benesty J, Sondhi M M, Huang Y A 2008 *Springer Handbook of Speech Processing* (Berlin: Springer)
[7] Schmidt-Nielsen A, Marsh E, Tardeli J, Gatewood P, Kreamer E, Tremain T, Cieri C and Wright J 2000 *Speech in Noisy Environments (SPINE) Evaluation Audio* (Philadelphia, PA: Linguistic Data Consortium)
[8] Kolokolov A.S., Lublinsky I.A. 2015 *Automation and Remote Control* 10 144–151
[9] Savchenko L V 2014 *Informatsionno-upravlyayushchiye sistemy* 1 23 (In Russian)
[10] Korsun O N and Poliev A V 2016 *Journal of Computer and Systems Sciences International* **55**(4) 609
[11] Oppenheim A V and Shafer R V 2010 *Discrete-Time Signal Processing* (Englewood Cliffs. NJ: Prentice-Hall)
[12] Ayvazyan S A, Bukhtshtaber V M, Yenyukov I S and Meshalkin L D 1989 *Prikladnaya statistika: Klassifikatsiya i snizheniye razmernosti* 607 (Moscow: Finansy i Statistika) (In Russian)
[13] Korsun O N and Gabdrakhmanov A S 2017 *Vestn. komp'yuternykh i informatsionnykh tekhnologiy* 1 10-15(In Russian)
[14] Luenberger D G, Yinyu Y 1984 *Linear and nonlinear programming* vol 2 (Reading, MA: Addison-wesley)