Attacking the problem of continuous speech segmentation into basic units

I.A. Andreev[1], A.I. Armer[1], N.A. Krasheninnikova[2], V.S. Moshkin[1]

[1]*Ulyanovsk State Technical University, Severny Venetz St., 32, 432027, Ulyanovsk, Russia*
[2]*Ulyanovsk State University, Lev Tolstoy St., 42, 432017, Ulyanovsk, Russia*

Abstract

The paper considers the algorithm of continuous speech segmentation into basic units, namely phonemes, certain combination of phonemes and pauses. The algorithm is based on speech signal transformation into a two-dimensional image, i.e. an autocorrelation portrait. To determine the boundaries of speech units the portraits of the analyzed signal are aligned with the model portraits of each speech unit. The authors apply the dynamic programming to find out the optimal distance between portraits.

Keywords: speech signal; segmentation; autocorrelation portrait; speech units; discrete dynamic programming

1. Introduction

At present, the algorithms for continuous speech segmentation into verbal units - phonemes, their combinations and pauses - are quite in demand. For example, this problem arises, while creating systems for research, processing, modeling and automatic speech recognition. To use such systems under different acoustic conditions, they should be subject to strict requirements for acoustic noise impedance and speech signal distortion. The article presents a method for determining the boundaries of speech pauses and speech units, which correspond to SAMPA + for the Russian language [1], [2]. The algorithms of speech signal transformation and processing used in the suggested method correspond to the strict requirements for acoustic noise impedance and speech signal distortion.

2. The subject of investigation

The problem of speech signal segmentation into its basic units is extremely complicated and challenging, and at present there is no simple solution for the general case. It is noted [3] that there are certain cases, for which exact segmentation is problematic. Different methods [3],[4],[5] are used for continuous speech signal segmentation. It is possible to distinguish the methods based on spectral analysis, trajectories of signal energy, energy logarithm, number of transitions through zero, and statistical parameters of speech units. The abovementioned methods give good results under favorable acoustic conditions, but the results deteriorate due to the presence of noise. Moreover, the time length of a speech signal varies from one pronunciation to another, which also makes its segmentation into basic units difficult. The authors suggest using the autocorrelation transformation [6],[7] of the speech signal into a two-dimensional image as well as the certain ways of image alignment in order to improve noise stability when determining the speech unit boundaries. The autocorrelation transformation has a number of characteristics, which make it somewhat noise-resistant [8]. Thus, the proposed method of speech signal segmentation is to be assumed to be less dependent on the current acoustic conditions, in which it was pronounced. Using discrete dynamic programming [9], when aligning two-dimensional speech signal images makes it possible to increase the stability of the method under consideration to the changes in the time length of speech units.

3. Algorithm for determining speech unit boundaries

1.1. General algorithm

The algorithm for determining speech unit boundaries is as follows: a speech signal containing a fragment of continuous speech analyzed for speech unit boundaries is represented in the form of digital readouts. The models of each speech unit are also represented as digital readouts. For benchmarking each example of the speech unit corresponding to SAMPA + is pronounced by the speaker, then the boundaries are defined by ear, and the speech unit becomes a model. By means of the autocorrelation transformation digital readouts of the analyzed continuous speech segment and the readouts of every model speech unit are transformed into particular two-dimensional images, which are called autocorrelation portraits (ACPs). For further alignment portraits of the analyzed speech segment and every model speech unit have the same line length.

Next, the portrait of the analyzed speech segment is aligned with all portraits of model speech units to determine the speech unit boundaries. For this purpose, the distance [10],[11] is calculated in the sliding window. The size of the window is equal to the number of lines in a corresponding speech unit portrait. During the calculation, the distance between the windows is optimized using the discrete dynamic programming. For each speech unit, a distance array along the portrait of the analyzed speech segment is determined. The distances corresponding to the same fragments of the analyzed speech segment portrait are compared with each other. As a result, speech unit portraits, which have the smallest distances, form the desired boundaries. If the smallest distance is obtained from the portraits of identical speech units, which follow one another, they are combined into the boundaries of one speech unit.

1.2. Autocorrelation portraits of speech signals

Since autocorrelation links are rather informative, i.e. they reflect speech signal features ACPs are unique for each speech unit. This provides good results in obtaining the speech unit boundaries for continuous speech. In [12] ACPs are modeled in the following way. Let $s(i)$ be the *i-th* readout of a digital speech signal; $s(i+k)$ is a readout spaced k readouts apart $s(i)$. Dependency factor of these readouts is expressed by a sample correlation coefficient:

$$R_s(k) = R[s(i), s(i+k)] = \frac{\mathrm{cov}[s(i), s(i+k)]}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} s^2(i) - m_{s(i)}^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N} s^2(i+k) - m_{s(i+k)}^2}},$$

$$\mathrm{cov}[s(i), s(i+k)] = \frac{1}{N}\sum_{i=1}^{N} s(i)s(i+k) - \left[\frac{1}{N}\sum_{i=1}^{N} s(i)\right]\frac{1}{N}\sum_{i=1}^{N} s(i+k), \tag{1}$$

where N is a number of readouts in the interval, in which the dependency is sought; $\mathrm{cov}[s(i), s(i+k)]$ is the sample covariance $s(i)$ and $s(i+k)$ when $i = 1..N$; $m_{s(i)}$ is a sample mean $s(i)$ when $i = 1..N$; $m_{s(i+k)}$ is a sample mean $s(i+k)$ when $i = 1..N$. Function determined by the sample correlation coefficient using (1) is an autocorrelation function (ACF) of a signal. While calculating ACF we perform the transformation of speech signal (SS) readouts $s(i) i = 1..M(M$ is the number of readouts in a speech signal) into a two-dimensional image. For this purpose, $s(i)$ is divided into intervals including $N < M$ readouts, then, in each $\mathrm{j-th}$ $(\mathrm{j} = 1, \mathrm{N}, 2\mathrm{N}, \dots, \mathrm{M} - 2\mathrm{N})$ interval the local signal maximum $\mathrm{i}_\mathrm{m}^\mathrm{j} = \max|\mathrm{s}|$ is sought. Let us assume that M is divisible by N evenly, otherwise the remaining final SS readouts are omitted. Then, using equation (1) we calculate the elements of the corresponding ACP line beginning with $i_m^j (j = 1, N, 2N, \dots, M - 2N)$ and generate ACP lines:

$$R\left[s\left(i_m^j\right), s\left(i_m^j + k\right)\right]_{j=1,N,2N,\dots,M-2N}^{k=1..N}, \qquad X(j,k) = R. \tag{2}$$

The two-dimensional image $X(j,k)$obtained from (2), where j is the line number, and k is the column number, is the ACP of a speech signal $s(i)$ dimensioned $N \times \left(\frac{M}{N} - 2\right)$, generated using SS local maxima. Note, that ACPs generated using local maxima are unique for each speech unit, and due to their link with SS local maxima they are less subject to geometrical distortions associated with speech variability. Figure 1 represents ACPs of speech units [“a], [o], [n`:], [f] (SAMPA+).

1.3. Alignment of autocorrelation portraits using discrete dynamic programming

Due to high degree of speech signal variability, autocorrelation portraits of one speech unit pronounced at different times differ from each other. Figure 2 shows ACPs of a speech unit “unstressed [a]”, one of them (a) was obtained from the pronunciation of the word «Вера» / “Vera”, and another (b) from the word «сопутствующие» / “soputstvujushhie”. It is obvious, that the portraits differ in the number of lines. Nevertheless, some lines of portrait a) can correspond to one line of portrait b).

The distance between the corresponding ACP lines is determined for the $i - th$ line of portrait X and the $j - th$ line of portrait Y using the following formula:

$$\rho_{i,j} = \sum_{k=1}^{N}\left(X(i,k) - Y(j,k)\right)^2. \tag{3}$$

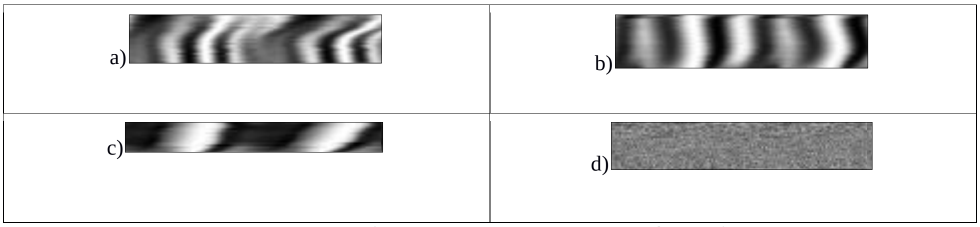

Fig. 1. ACPs of speech units a) [“a], b) [o], c) [n`:], d) [f].

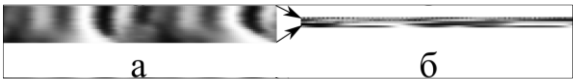

Fig. 2. ACPs of a speech unit “unstressed [a]”: a) model, b) as a part of the word «сопутствующие» / “soputstvujushhie”.

To determine the measure of ACP concordance the discrete dynamic programming [9] is applied. It allows to minimize the functional $\rho = \min_{\Omega}\sqrt{\sum \rho_{i,j}}$, which characterizes ACPs identity. Set Ω predetermines the permitted correspondences of the portrait lines, which are obtained on the basis of the following rules. 1. The number of lines in ACPs can differ. 2. Any line of one particular ACP cannot correspond to the line of another one spaced from the previous corresponding line more than c lines apart. 3. The order of line correspondence is preserved, i.e. if the i-th line of one ACP corresponds to the j-th line of the other one, then the $(i+1)$-th line cannot correspond to $j - l, l = 1,2, \dots$ 4. The total distance between ACP pronunciations of the same speech units formed from the distances between the corresponding lines according to the second metrics rule should be minimal according to rules 1)-3).

To determine the measure of speech signal ACP correspondence (in a two-dimensional sliding window) to speech unit ACP the following algorithm is obtained. Matrix D containing $m \times m$ elements is created, where m is the number of CP lines in a sliding window X; the number of speech unit Y ACP lines is the same. For example, let $c = 3$. At first, the distances between $Y(1)$ and $X(1), X(2), X(3)$ are found, then these distances are stored in D

$$D_{1,i} = \rho\big(Y(1), X(i)\big), i = 1..3. \tag{4}$$

Then, distances between $Y(2)$ and $X(1), X(2), X(3), X(4), X(5)$ are found. The position of the line $Y(1)$ is taken into account, i.e. if $Y(1)$ corresponds to $X(2)$, then $Y(2)$ can be compared only with $X(2), X(3), X(4)$. Each time it is necessary to remember portrait X line number, and fill in the matrix $D_{2,i} = D_{1,i} + \rho\big(Y(2), X(j)\big), j = i..i+2$. Besides, each element from D due to intersection of possible line positions can be filled in several times. In such a case, the minimum value (Figure 3) is preserved:

$$D_{k,i} = min\big[D_{k,j}, D_{k-1,j} + \rho\big(Y(k), X(j)\big)\big], j = i..i+2. \tag{5}$$

During the next stages, all the remaining matrix D elements are found using formula (5), at each stage i changes from 1 to $I + 2$, where I is the maximum value of i at the previous stagee. For the first stage $I = 1$. The algorithm is stopped when matrix D is completely filled. The minimal element from the m-th line and the m-th column of the matrix corresponds to the minimal distance between X and Y.

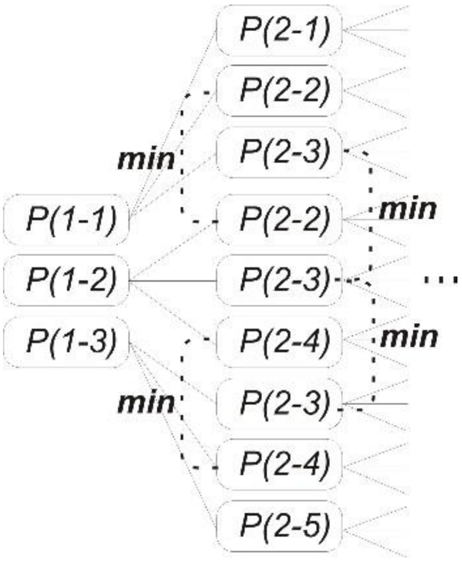

Fig. 3. Distribution of compared ACP lines. P(i-j) is the distance between the i-th line of one ACP and the j-th line of another one. Mark *min* shows that from all possible identical comparisons at different stages of programming the comparison with the minimal **distance** is chosen.

4. Experiments

The suggested algorithm for determining speech unit boundaries in continuous speech was tested experimentally. Figure 4 shows the speech unit boundaries in the utterance containing the pronunciation of the word «основного» / “osnovnogo”. For example, the interval of speech unit [a] pronunciation, which starts the word «основного» / “osnovnogo”, was correctly defined within the range from 800 to 4800 speech signal digital readouts, speech unit [s] – in the range from 2400 to 5600 readouts, speech unit [n] – in the range from 5600 to 9200 readouts, speech unit [a] – in the range from 9600 to 11200 readouts, speech unit [v] – in the range from 11200 to 16000 readouts, speech unit [n] – in the range from 16000 to 17200 readouts, speech unit [“o] in the range from 17200 to 26400 readouts, speech unit [v] – in the range from 26400 to 28000 readouts and the last of the analyzed speech signal unit [a] – in the range from 28000 and up to the end of the signal.

Comparison with expert borders was not made. However, visual comparison of the determined boundaries with the real ones shows their closeness. Experiments show the practical applicability of the algorithm for determining the speech unit boundaries in continuous speech.

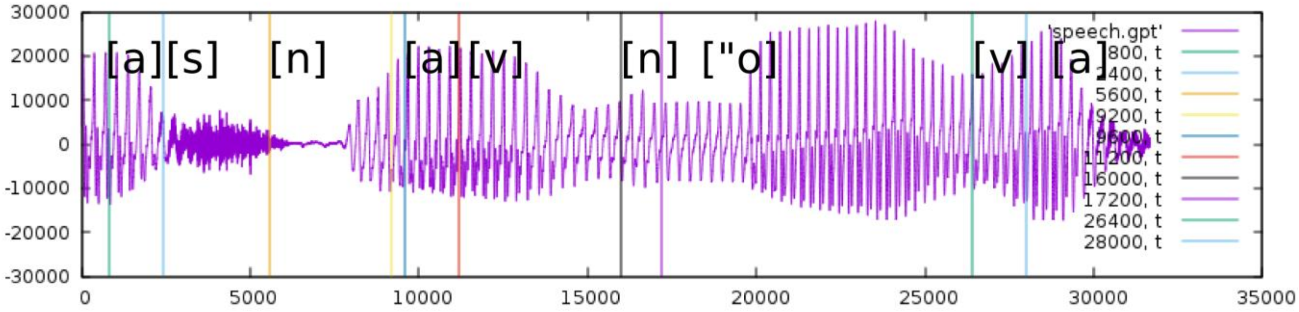

Fig. 4. Speech unit boundaries in continuous speech containing the pronunciation of the word «основного» / “osnovnogo”.

5. Conclusion

The determined speech unit boundaries are to be used for a more detailed analysis of the speech signal in order to identify the speech units. In order to solve this problem the authors also want to transform speech signals into ACPs. However, the parameters of transformation into ACPs and the method of portrait alignment will be different.

Acknowledgements

The work was supported by grants 16-48-732046 and 16-48-730305 from the Russian Foundation for Basic Research.

References

[1] Galounov VI, Heuvel H, Kochanina JL, Ostroukhov AV, Tropf H, Vorontsova AV. Speech Database for the Russian Launguige. Proceedings of international workshop SPEECOM 1998.

[2] Michael P, Rasanen O, Thiollière R, Dupoux E. Improving Phoneme Segmentation With Recurrent Neural Networks. Computation and Language, 2016, preprint:1608.00508.

[3] Rabiner LR, Schafer RV. Digital processing of speech signals. Edited by M.V. Nazarov and Yu.N. Prokhorov. Moscow: Radio i svyaz', 1981; 496 p. (in Russian)

[4] Goldenthal W. Statistical Trajectory Models for Phonetic Recognition. PhD thesis. M.I.T., 1994; 170 p.

[5] Ostendorf M, Roukos SA. A stochastic segment model for phoneme-based continuous speech recognition. IEEE Transaction on Accoustics, Speech, and Signal Processing 1989; 37(12): 1857–1869.

[6] Therrien C, Tummala M. Probability and Random Processes for Electrical and Computer Engineers. CRC Press, 2012; 287 p.

[7] Amirgaliyev Y, Mussabayev T. The speech signal segmentation algorithm using pitch synchronous analysis. Open Comput. Sci. 2017; 7: 1–8.

[8] Krasheninnikov VR, Armer AI, Krasheninnikova NA, Kuznetsov VV, Khvostov AV. Some problems connected with speech command recognition on the background of intense noise. Infokommunikatsionnye tekhnologii. Samara 2008; 1: 72–75. (in Russian)

[9] Bellman R. Dynamic programming. Moscow: IL, 1960; 400 p. (in Russian)

[10] Krasheninnikov VR, Armer AI, Kuznetsov VV. Autocorrelated Images and Search for Distance between them in Speech Commands Recognition. Pattern Recognition and Image Analysis. 2008; 18(4): 663–666.

[11] Greibus M. Rule Based Speech Signal Segmentation. Journal of telecommunications and information technology 2010; 4: 37–43.

[12] Krasheninnikov VR, Armer AI, Krasheninnikova NA, Khvostov AV. Speech command recognition on the background of intense noise using autocorrelated portraits. Naukojomkie tehnologii 2007; 8(9): 65–76. (in Russian)