Modification of entropy-based algorithm for determining words boundaries in noisy conditions

A S Karpov, V I Drozdova, G V Shagrova and M G Romanenko

Department of Information Systems and Technologies, North-Caucasus Federal University, 1, Pushkin Street, Stavropol 355017, Russia

E-mail: victoria_drozdova@rambler.ru

Abstract. A modification of the algorithm for finding word boundaries in continuous speech based on an analysis of the entropy of a speech signal is described. At the preprocessing stage, the signal is divided into frames; the entropy value is calculated for each frame and compared with the threshold value of the entropy \( \gamma \). The algorithm is carried out in two stages, first a rough and then more precise definition of the word boundaries is made. At each stage, its own minimum distance between words is used. A modification is proposed, which consists in the fact that the entropy threshold \( \gamma \) is determined on the basis of the calculated SNR coefficient taking into account the experimental selection of signal parameters at the stage of precise determination of word boundaries in a noisy speech signal. Checking the functionality of the modified algorithm on noisy signals has shown that taking into account the effect of SNR on the selection of signal parameters increases the accuracy of word boundary determination.

1. Introduction

Automation of speech recognition systems is relevant both for traditional applications and for new breakthrough technologies. Automatic speech recognition began at the end of the 60s on the basis of the dynamic programming method for individual commands [1], followed by Markov chains, Fourier transform and wavelet transform, Format analysis and signal decomposition into empirical modes (EM) [2,3], cepstral analysis [3], correlation analysis, randomized algorithms of stochastic approximation [4], algorithms based on the calculation of speech signal entropy [5,6] and neural networks [7,8]. New and rapidly developing areas of application of speech recognition systems include: telephony: automation of processing of incoming and outgoing calls by creating voice self-service systems; Voice control system interface "Smart Home"; voice control of household appliances and robots; voice input in computer games and applications; voice control in the car; medical diagnostics of borderline human mental disorders [9] and a method for detecting psychoemotional human disorders based on the decomposition of speech signals into empirical modes and their format analysis [2,3]; systems that automatically generate dictated text [1].

Continuous speech consists of discrete components from phoneme to utterance, which are analyzed at various stages of signal processing. In particular, at the preprocessing stage, after filtering and noise suppression, informative sections of speech signals are extracted. The definition of word boundaries in the speech stream is an inherent part of the speech recognition process. It is at this stage that fragments that do not carry speech information are excluded from the signal, which increases the speed of the speech recognition system. The implementation of algorithms for determining word boundaries based on statistical processing of signals [4], the signal analysis in time domain, where the envelope of the
signal is calculated and compared with previous created pattern stored in memory [10], entropy calculation [11] does not require the use of large computational resources, which can be useful for servicing mobile devices [6], recognition of keywords and individual commands, as well as the use of modern intelligent methods at the preprocessing stage [8]. Methods based on the calculation of speech signal entropy retain their effectiveness even under conditions of sharp fluctuations in signal amplitude caused by noise or the acoustic characteristics of the pronunciation of certain sounds [6, 11]. The approach considered in this paper makes it possible to use the results of analysis of the entropy of a speech signal to search for word boundaries. The aim of the work is to improve the accuracy of determining the boundaries of words in a speech signal by modifying the algorithm described previously by the authors in [12].

2 Materials and methods
To determine the word boundaries, an algorithm based on the calculation of the entropy of the speech signal is used, according to which in the preprocessing step, the following is performed:
1) normalization of digitized speech signal at which amplitude values range from 1 to -1;
2) dividing the normalized signal into frames by 25 milliseconds, overlapping by 25-50%.

Then, for each frame of the analyzed speech signal, the entropy value is calculated according to the technique described in [12]. As a result, the histogram of values of entropy of a recognizable speech signal for each frame turns out (1 ≤ j ≤ m). This histogram is called the entropy profile ξ:

\[ \xi = [H_1, H_2, ..., H_m]. \] (1)

The obtained values of the entropy of the speech signal are compared with the value of the threshold of entropy γ, which is calculated by the formula:

\[ \gamma = min(\xi) + (max(\xi) - min(\xi)) \cdot k \] (2)

where k is the coefficient that was selected experimentally.

Any value of \( H_j \) equal to or greater than the threshold of entropy is considered a speech, and all that is less is silence or noise. However, as a result of the comparison, some frames containing speech may be discarded as non-speech data. To avoid the erroneous definition of word boundaries in a speech signal, the concepts of the minimum word length (h) and the minimum distance between words (δ) [5] are entered. Both of these values are characterized by the number of frames. With h and δ in mind, each recognized speech segment \( \lambda \) must have a certain minimum length, which is indicated as a constant.

That is, if \( \lambda_i < h \) and the distance between adjacent segments \( d_{ij} > \delta \), then this segment is discarded as a segment that does not contain speech information. Two segments of the analyzed speech signal, defined as speech, are combined into one if the distance between them (\( d_i \)) is less than the specified number of frames. The described algorithm is carried out in 2 stages, first a rough and then more precise definition of word boundaries is performed. At each stage, its own minimum distance between words (δ) is used. The presented algorithm copes well with determining the word boundaries of continuous speech recorded under ideal conditions (without distortion and noise) [12]. In this work, this algorithm has been tested for noisy audio signals, which are a recording of continuous speech, as well as modification, which allows to more accurately determining the boundaries of words in the noisy speech signal.

To estimate the influence of noise on the speech signal, the SNR indicator was used, which is expressed in decibels (dB) and can be calculated.

\[ SNR = \frac{P_{signal}}{P_{noise}} = \left( \frac{A_{signal}}{A_{noise}} \right)^2 \] (3)

where P is the average signal strength (noise); A - RMS value of signal (noise) amplitude.

The higher the SNR, the less noise affects the characteristics of the speech signal [13, 14, 15].

\[ SNR(dB) = 10 \log_{10} \left( \frac{P_{signal}}{P_{noise}} \right) = 20 \log_{10} \left( \frac{A_{signal}}{A_{noise}} \right) \] (4)
In telephone systems, the SNR is about 30 dB \([4]\), so it was assumed that the operation of the algorithm should determine the word boundaries in a speech signal for which the SNR is 30 dB or more.

Taking into account the estimate of signal power and noise depending on the SNR value of the signal under study, the entropy threshold is calculated as follows:

\[
\gamma = \begin{cases} 
\text{SNR}(\text{Db}) > 30, & \min(\xi) + (\max(\xi) - \min(\xi)) \cdot 0.9 \\
\text{SNR}(\text{Db}) < 30, & \min(\xi) + (\max(\xi) - \min(\xi)) \cdot 0.95 
\end{cases}
\]  

(5)

To test the algorithm’s performance under noisy conditions, white noise of various powers was artificially superimposed on the recording under study. Checking the original algorithm on the studied sound signals showed that it does not allow determining the boundaries of all words in a noisy signal. Therefore, a modification of the algorithm described in \([12]\) was proposed, which consists in the fact that the entropy threshold \(\gamma\) is determined at the stage of exact determination of word boundaries in a noisy speech signal based on the calculated value of the SNR coefficient taking into account the experimentally selected coefficient \(k\).

Thus, a modified algorithm for determining word boundaries, as well as described previously in \([12]\), is performed in two stages. The first stage is reduced to the following operations:

1. Pretreatment. Dividing the signal into frames.
2. Rough definition of word boundaries.

The second stage of the modified algorithm is characterized in that the exact definition of word boundaries is made taking into account SNR values and is reduced to the following operations:

1. Calculation of threshold values of entropy taking into account SNR according to (5).
2. Comparison of the values of entropy with the threshold of entropy

\[
\xi = \begin{cases} 
H_j, & H_j \geq \gamma \\
0, & H_j < \gamma
\end{cases}
\]

3. Checking the criteria

\[
\left\{ \begin{array}{l}
\lambda_i < h, \quad d_{ij} > \delta \\
\left( \lambda_i, \lambda_j \right) > h, \quad d_{ij} < \delta
\end{array} \right.
\]

4. Formation of groups of frames corresponding to individual words.

3. Results

The results of the original and modified algorithms are analyzed using the example of the phrase \(\lambda:\) “Last spring was very warm” (in Russian). The figure 1 shows the result of the original algorithm obtained by determining the word boundaries of continuous speech recorded under ideal conditions (without distortion and noise). At the second stage of the algorithm, only those groups of frames that were formed after the first stage were considered. For them, the parameters were used: \(k = 0.75\), and the minimum distance between words was 3 frames, since the minimum duration of sounding of phonemes is approximately 50 ms. As shown in Figure 1, the original algorithm identified the boundaries of all five words.

The experimental results obtained on sound signals of continuous speech under conditions of noise of different power are shown in Table 1. The table shows the number of selected frame groups corresponding to the number of words in the analyzed speech signal for different values of the SNR and \(k\) parameters. The table shows that for \(k = 0.9\), the algorithm provides an accurate definition of word boundaries in the signal for which SNR = 30 dB or more. An increase in the coefficient \((k)\) to 0.95 facilitated the accurate determination of word boundaries in a speech signal for which SNR = 22 dB, but negatively affected the operation of the algorithm with signals for which SNR = 35 dB or more. Thus, a change in SNR affects the operation of an algorithm based on determining the entropy of a speech signal.
Figure 1 - The result of the algorithm for the phrase "Last spring was very warm"

Table 1. The number of selected groups of frames

|       | SNR =22 dB | SNR =30 dB | SNR =35 dB | SNR =40 dB | SNR =50 dB |
|-------|------------|------------|------------|------------|------------|
| k = 0,75 | 6          | 4          | 6          | 6          | 6          |
| k = 0,8 | 6          | 6          | 3          | 4          | 3          |
| k = 0,85 | 8          | 7          | 6          | 6          | 7          |
| k = 0,9 | 4          | 5          | 5          | 5          | 5          |
| k = 0,95 | 5          | 5          | 7          | 6          | 7          |

Figure 2a presents the results of the initial algorithm in which the entropy threshold at the second stage was calculated by formula (2), and a modified algorithm in which the entropy threshold at the second stage was determined according to (5). From the results presented in figure 2, it follows that the original algorithm did not cope with the task of determining word boundaries, highlighting four groups of frames. A modified algorithm accurately determined the boundaries of all five words. By calculating the SNR for the input speech signal, it is possible to vary the coefficient (k) depending on the degree of noise of the recognized speech stream. The modified algorithm allocated 5 groups of frames in the conditions of noise. Therefore, to more accurately determine the word boundaries in a noisy signal, it is proposed to use formula (5) to determine the entropy threshold.

4. Discussion
As a result of the analysis of the data of a computational experiment, it was found that the original algorithm allows you to accurately determine the word boundaries in continuous speech recorded under ideal conditions. When applying noise of different power to this speech signal, this algorithm does not allow determining the boundaries of all words in the analyzed message. Therefore, to more accurately determine the word boundaries in a noisy sound signal, it is advisable to use an approach in which the entropy threshold at the second stage of the algorithm is calculated based on the coefficient k, which
must be selected taking into account the values of the SNR parameter. This approach provides a more accurate definition of the boundaries of speech signal words in noise conditions. However, for noisy signals it is advisable to check the values of the variable coefficient on a more representative sample.

![Figure 2. The result of the original (a) and modified (b) algorithms in a noisy environment](image)

5. Conclusion
An algorithm is presented that allows you to select individual words in continuous speech based on the definition of the entropy of a speech signal. The relative simplicity and low cost of implementing the algorithm can find application at the stages of signal preprocessing. A modified algorithm for determining word boundaries in continuous speech based on the calculation of the entropy of a speech signal is presented. The process of determining word boundaries is carried out in two stages. At the first
stage, a rough selection of large groups of frames containing voice information is performed. At the second stage, more detailed segmentation of speech fragments obtained at the first stage takes into account the SNR value.

References

[1] Tampel I  B  2015 Journal of Information Technologies, Mechanics and Optics 15:6  957 (In Russian) doi:10.17586/2226-1494-2015-15-6-957-968
[2] Alimuradov A K, Tychkov A Yu, Kuzmin A V, Churakov P P, Agyekin A V and Vishnevskaya G V  2017 Proc.Int.Conf. Open Innovations Association FRUCT (Helsinki/ IEEE) p26
[3] Alimuradov A K, Tychkov A Yu, Agyekin A V and Churakov P P  2018 Proc.Int.Conf on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) (Yekaterinburg: Russia / IEEE) p 128
[4] Granichin O N, Shalymov D S, Avros R and Volkovich Z  2011 Autom. Remote Control, 72:4  754
[5] Obin N and Liuni M  2012 IEEE workshop on Spoken Language Technology, United States
[6] Mustafa M K  2016 On-device Mobile Speech Recognition – Dr.Ph Thesis Nottingham Trent University (Nottingham:UK) p 149
[7] Romanenko A N  2018 Scientific and Technical Journal of Information Technologies, Mechanics and Optics 18 : 2, 350 (in Russian) doi: 10.17586/2226-1494-2018-18-2-350-352
[8] Nazarov A V , Yakimov V L and Avdeev V A  2015 Information and Control Systems (http://www.i-us.ru): 75 :2.57 (In Russian) doi:10.15217/issn1684-8853.2015.2.57
[9] Alimuradov A K, Tychkov A Yu, Churakov P P and Agyekin A V  2017 University proceedings. Volga region Engineering sciences. Computer science, computer engineering and control 43 : 3 4 (in Russian) doi: 10.21685/2072-3059-2017-3-1
[10] Raczynski M  2018 Proc.Int. Interdisciplinary PhD Workshop (IIPhDW)(Swinoujście:Poland) p27 doi: 10.1109/IIPHDW.2018.8388238
[11] Kochetkov A V and Fedotov P V  2015 Internet journal “Naukovedenie” (http://naukovedenie.ru)/, 7(6) p1 (in Russian) doi: 10.15862/72TVN615
[12] Karpov A S, Drozdova V I, Shagrova G V. and Shevchenko AV  2018 CEUR Workshop Proceeding 2254  162
[13] Recommendation ITU-T G.160  2008 Objective measures for the characterization of the basic functioning of noise reduction algorithms  (Geneva: Switzerland) p 72
[14] Stolbov M B and Quan Trong  2020 Journal of Information Technologies, Mechanics and Optics 20(1) 147 (in Russian). doi: 10.17586/2226-1494-2020-20-1-147-154.
[15] Kim C and Stern R M  2008 Proc. Proc.Int.Conf. INTERSPEECH (Brisbane: Australia) P 2598