Entropy-based detection of the words boundaries of continuous speech

Andrey S. Karpov
Information Systems Technologies Dept.
Stavropol, NCFU
andrey_revol125@mail.ru

Victoria I. Drozdova
Information Systems Technologies Dept.
Stavropol, NCFU
victoria_drozdova@rambler.ru

Galina V. Shagrova
Information Systems Technologies Dept.
Stavropol, NCFU
g_shagrova@mail.ru

Aleksey V. Shevchenko
Information Systems Technologies Dept.
Stavropol, NCFU
luckyleo769@mail.ru

Abstract

An algorithm for finding the word boundaries in a merged speech is proposed on the basis of a method using the definition of the entropy of a speech signal. The difference between the proposed algorithm and the known ones is the comparison of the speech signal entropy value with the entropy threshold in two stages. The work of the known and proposed algorithm is compared.

1 Introduction

Automatic speech recognition, especially in noisy environments, is a complex task. The most important step in automatic speech recognition is the correct definition of word boundaries in the speech stream. Even a slight improvement at the stage of delineating the boundaries of words significantly affects the performance of the entire speech recognition system.

To recognize isolated words, this problem reduces to determining the correct word boundary. For confluent speech, this task is much more difficult, since the speech signal is a continuous stream without any speech pauses.

The most promising approach involves the use of speech signal entropy to search for word boundaries [Koc15, Naz15]. The main feature of the method using the entropy value of the speech signal is low sensitivity to changes in the amplitude of the speech signal, which leads to the preservation of more detailed information contained in the speech stream. For a speech recognition system to be effective, it must work satisfactorily in an environment where the incoming voice signal is noisy. Even in the presence of small noise in the analyzed speech signal, the approach using the entropy value provides high accuracy of the definition of word boundaries.

2 Description of the object and methods of research

Defining the boundaries of words in a speech signal is a key aspect of the human speech recognition task, as at this stage speech data is separated from unnecessary noise and speech artifacts (a cough, speech harmonics,

In: Marco Schaerf, Massimo Mecella, Drozdova Viktoria Igorevna, Kalmykov Igor Anatolievich (eds.): Proceedings of REMS 2018 – Russian Federation & Europe Multidisciplinary Symposium on Computer Science and ICT, Stavropol – Dombay, Russia, 15–20 October 2018, published at http://ceur-ws.org

microphone echo, etc.).

The use of the method based on the value of the entropy of the speech signal gives high indicators of the definition of word boundaries for the task of recognizing isolated commands [Alu14, Alu16, Boz11, Wah02].

The essence of the method is that the input voice data is preliminarily processed using a bandpass filter. This filter removes the constant and low-frequency components of the background, as well as high-frequency noise and speech harmonics, arising from the spectral properties of the voice path. The pre-processed speech is normalized so that the amplitude values of the signal lie in the range from 1 to -1. Then, the normalized signal is divided into frames of approximately 25 milliseconds of speech. To avoid loss of information, these frames have an overlap of 25 - 50%.

Then the entropy value in each frame of the analyzed sound sequence is calculated:

$$H = -\sum_{i=1}^{N} p_i ln(p_i) \tag{1}$$

where $H_j(j = 1, 2, \ldots, m)$ – the value of entropy of the j-th frame, m – the number of frames;;
p_i – the probability of i-th signal count, in j-th frame;
N – the number of counts within the frame.

As a result, the entropy profile ξ is determined for the incoming speech signal, which is a histogram of the entropy values of all the frames, the recognizable fragment of speech:

$$\xi = [H_1, H_2, H_m] \tag{2}$$

In the case of recognizing isolated instructions, the entropy profile of the signal is used to calculate the entropy threshold γ.

$$\gamma = \frac{max(\xi) - min(\xi)}{2} + \mu min(\xi); \mu > 0 \tag{3}$$

where μ – the noise ratio, which is selected experimentally [Boz11].

However, for example, in [Alu14, Alu16] the value of the entropy threshold is not calculated but is taken equal to a constant value: γ=0,1. But this approach does not give good results for cases of a noisy signal.

After determining the threshold, the value of the entropy of each frame H_j is compared with the entropy threshold γ. Any value equal to or greater than the entropy threshold is considered a speech and all that is less is silence or noise.

$$\xi = \begin{cases} H_j, & H_j \geq \gamma \\ 0, & H_j < \gamma \end{cases}, \tag{4}$$

However, due to the vocal characteristics of the speech signal, the entropy index may be too small in the area of the recognizable speech signal that carries the information. Or, conversely, because of instantaneous noise, a signal segment that does not carry speech data is recognized as speech [Naz15]. In order to avoid the erroneous definition of the word boundaries in the speech signal, the concepts of the minimum word length (k) and the minimum distance between words (δ) [Alu14, Alu16, Obi12]. Both these quantities are measured in the number of frames.

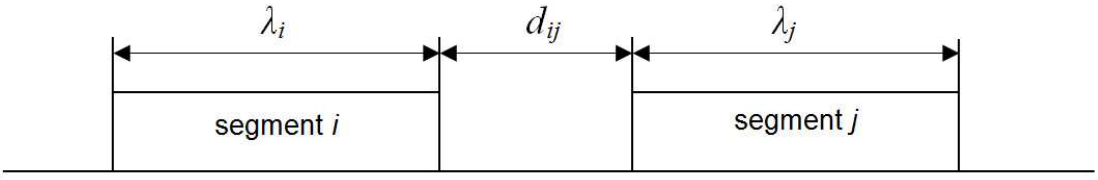


Figure 1: Communication between speech segments

The first criterion is that each recognized speech segment (λ_i, λ_j) must have a certain minimum length, which is indicated as a constant. That is $\lambda_i < k$ and $d_{ij} > \delta$, the i-th segment is discarded as a segment that does not contain voice information. Also, if $\lambda_j < k$ and $d_{ij} > \delta$, the j-th segment is discarded.

The second criterion is based on the minimum distance between words It consists in the fact that two segments of the analyzed speech signal, defined as speech, are combined into one if the distance between them (d_{ij}) is less than the specified number of frames. This means that if (λ_i or λ_j) $> k$ and $d_{ij} < \delta$, then the two segments are combined into one.

This approach gives a high result of detecting the boundaries of isolated words. In order to use this approach in the recognition of the continuous speech, an algorithm is proposed, according to which the entropy threshold was determined by the formula (5):

$$\gamma = min(\xi) + (max(\xi) - min(\xi)) \cdot k \tag{5}$$

where k – the coefficient that was selected experimentally, the word boundaries were determined in two stages. At each stage, the minimum distance between words (d_{ij}) was used.

The result of the proposed algorithm is given in the work by the example of separating the boundaries of the words of the merged speech, which is a speech signal containing the phrase "Dear passengers, please keep calm, the train will soon leave" pronounced in a woman's voice.

The analyzed phrase consisting of eight words was recorded with a sampling frequency of 22kHz, the number of channels 2 (stereo) and 16 bits. The duration of the speech signal was 6,583 seconds. The boundaries of the words of this phrase were in two stages.

The minimum distance between words in the first stage was 12 frames, and k = 0,9. This means that all frame groups defined as speech, but less than 12 frames in length, are discarded as non-verbal data. The results of the first stage of the algorithm are shown in Figure 2.

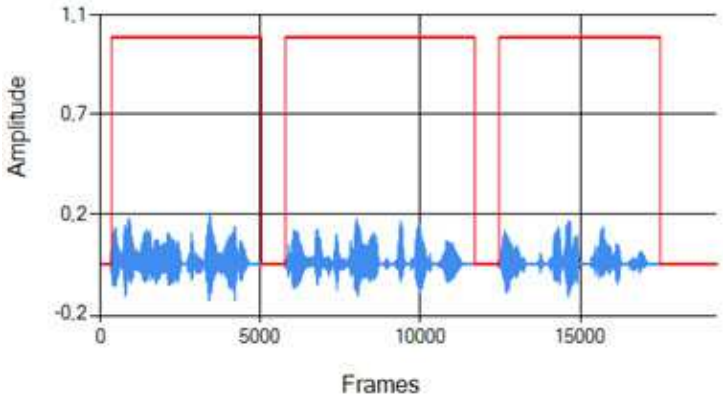


Figure 2: Boundaries of frames with voice information, obtained at the first stage of the algorithm execution

As shown in Figure 2, as a result of the first stage of the algorithm, three large groups of frames carrying the voice information were formed. At the second stage, only those frame groups that were formed after the first stage were considered. For them, the minimum distance between words was 3 frames, and k = 0,75. Let us consider the work of the second stage of the algorithm for frame groups formed after the first stage (Figure 3).

As can be seen from Figure 3, in the second stage, the algorithm divided the first large group of frames into two smaller ones, which are separate words. Similarly, both the second and third large groups were divided into three smaller ones. As a result, the boundaries of all eight words were found.

An example of a comparison of the work of the known and proposed algorithms is given for the speech signal, which is the phrase "Today is good weather". The analyzed phrase is pronounced by a man and recorded with a sampling frequency of 16 kHz, the number of channels 1 (mono) and 8 bits. The duration of this phrase was 2,535 seconds. The results are shown in Figure 4.

As can be seen from Figure 4, the known algorithm defines the entire phrase as a group of frames that carry information. Whereas the proposed algorithm determines the boundaries of all three words quite accurately.

3 Summary

A new algorithm for determining the boundaries of words in a merged speech is proposed, which differs from the known, based on the calculation of the entropy value of a speech signal, in that the process of separating the boundaries of words is performed in two stages. At the first stage, a rough selection of large groups of frames containing verbal information is carried out. At the second stage, there is a more detailed segmentation of the speech fragments obtained in the first stage.

Due to the use of the method, based on the definition of the entropy of the speech signal in speech recognition systems, much higher recognition rates of the word boundaries can be achieved, both in isolated and in the combined speech.

References

[Alu14] D.Yu Alunov. On Methods for Estimation of the Signal Parameters. *Current Problems of Science and Education No. 6*, 2014.

[Alu16] D.Yu. Alunov, E.S. Sergeev, P.V. Pigachev, A.N. Mytnikov. Implementation of the algorithm for processing and recognizing speech. *Modern high technology No. 3-2*, pp. 225- 230, 2016.

[Boz11] A.S. Bozhdai, P.A. Gudkov , A.A. Gudkov. Embedded identification system by voice biometric indicators . *Open Education No 2-2*, 2011.

[Koc15] A.V. Kochetkov, P.V. Fedotov. About various meanings of the concept "entropy". *Internet-journal Naukovedenie, Vol. 6*, 2015.

[Naz15] A.V. Nazarov, V.L. Yakimov, V.F. Avdeev. The algorithm for maximizing the entropy of the training sample and its use in the synthesis of forecast models for discrete states of nonlinear dynamical systems. *Scientific Journal "Information Control Systems": Issue 2*, St. Petersburg, 2015.

[Obi12] N. Obin, M. Liuni. On the generalization of shannon entropy for speech recognition. *IEEE workshop on Spoken Language Technology*, United States, 2012.

[Wah02] K. Waheed, K. Weaver, F.M. Salam. A robust algorithm for detecting speech segments using an entropic contrast. *Midwest Symposium on Circuits and Systems 3*, 2002.

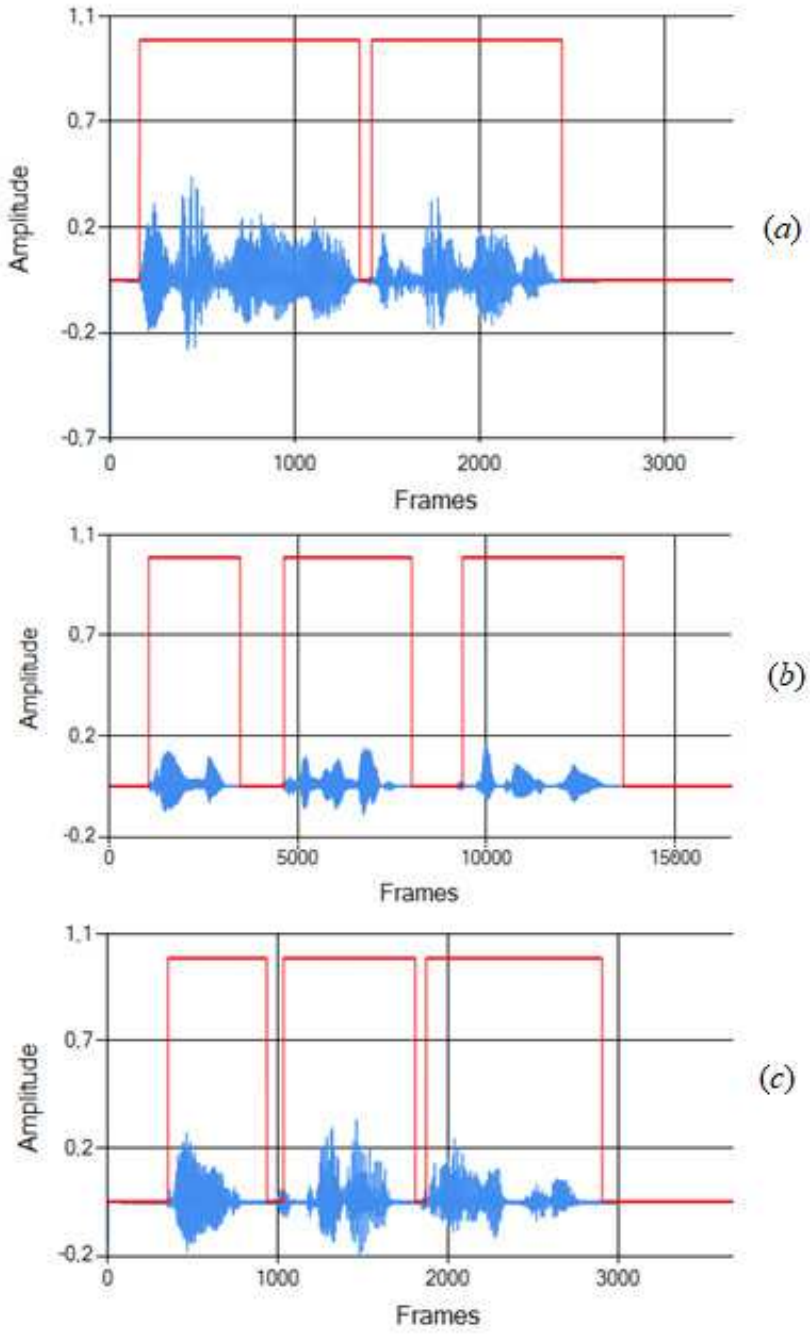
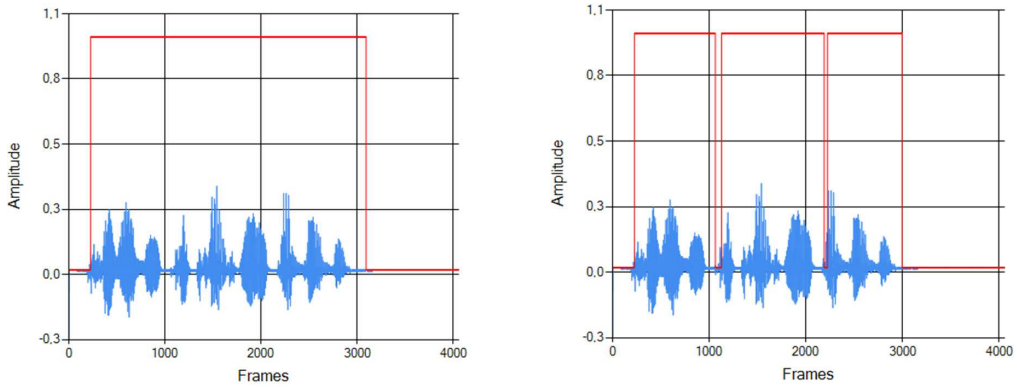


Figure 3: The result of the second stage of the algorithm for the first (a), second (b) and third (c) frame groups that formed after the first stage of the algorithm



(a) (b)

Figure 4: The result of the known (a) and proposed (b) algorithm for the phrase "Today is fine weather"