# New Aggregative Algorithms for Robust Speech Signal Segmentation

**Krassovitsky*, R. Mussabayev***

*Institute of Information and Computational Technologies, Pushkin str. 125, Almaty, Kazakhstan*

**Corresponding author's e-mails: akrassovitskiy@gmail.com, rmusab@gmail.com*

**Abstract**

Speech signals in natural language and machine learning methods used for analysis and recognition have attracted a lot of attention for many years as an object of research. The purpose of our work is to develop algorithms with cheap computation costs that allow extracting helpful information about the structure of a raw speech signal in an unknown language using unsupervised methods.

*Keywords:* ASR, speech features, FIR-analysis, clustering, machine learning

## 1 Introduction

We present new robust algorithms for multilevel speech signal segmentation as a part of the possible framework for automatic classification of received speech signal[1]. The obtained algorithms have a sufficient degree of robustness and allow for multilevel segmentation of the analyzed speech signal. The resulting multilevel segments of the speech signal are subjected to further classification in the context of different classes. The segmentation is performed at various scale levels that include the acoustic, and microwave levels.

## 2 Algorithmic approaches

Various methods and approaches for acoustic segmentation of a speech signal have been developed and implemented in software [2,3,4,5]. All developed methods were experimentally tested on the generated dataset. In the process of experimental verification, the effectiveness of the proposed algorithms for automatically determining the boundaries of acoustically homogeneous segments was evaluated. The measure of correspondence between the automatically obtained boundaries of segments and the manual labeling available in the dataset was numerically evaluated. If the studied algorithm had its parameters, their parametric identification was carried out to maximize the selected quality criterion to measure the correspondence between automatic and manual labeling. In the course of the experiments, it was important to evaluate the ratio of the speed and quality of the generated labeling, as well as the simplicity of the algorithm itself, both in terms of the simplicity of the algorithm structure itself and in terms of minimizing their customizable parameters.

The task of the developed algorithm is to identify local acoustically homogeneous sequences of basic frames of the analyzed speech signal. This algorithm is based on the analysis of a local homogeneity measure of the speech signal spectral characteristics. At the initial stage, the selected speech signal is subjected to pre-processing, including normalization, framing, and parameterization, i.e., calculation of basic informative features for selected frames and their subsequent higher-level analysis. During normalization and parameterization, the original speech signal is processed, and then it is automatically framed, i.e., it is divided into relatively short, evenly spaced, and overlapping frames (short sections containing the original speech signal). Before parameterization, each frame is weighted by a specialized window function in order to minimize the effect of a jagged or distorted spectrum. Thus, at the initial stage of acoustic segmentation, the basic units of analysis are parametrized frames presented as an ordered set of their compressed spectral characteristics. Each frame is already described not by a section of the original speech signal but by the corresponding vector of features that characterize its compressed spectral characteristic in the form of a finite set of spectral coefficients and their derivatives. Each frame is specified by 80 parameters, including the base spectral coefficients and the first and the second derivatives.

Thus, based on the results of framing and parameterization of the speech signal, we obtain a temporal sequence of vectors of features of the speech signal. And the very task of acoustic segmentation of a speech signal is reduced to finding local clusters of feature vectors in their resulting time sequence as they analyze the local commonality of their set of features. Local uniformity means the presence of similar spectral or acoustic characteristics in adjacent segments or in a group of adjacent segments.

### 2.1 ACOUSTIC SPEECH SIGNAL SEGMENTATION

Essential requirements for speed, accuracy, and robustness are imposed on the efficiency of the developed algorithm for acoustic segmentation of a speech signal [6,7]. These relatively high requirements are because this algorithm is one of the most fundamental algorithms and its efficiency directly affects the efficiency of other higher-level segmentation algorithms. The acoustic level segmentation

algorithm should be relatively fast, and it is desirable to have a linear dependence on the amount of processed speech data. This algorithm should analyze the local context of the speech signal exclusively, regardless of the global context; that is, segmentation should be carried out mainly at the local acoustic level without considering the global structure of this analyzed speech signal. The level of the local segmentation of the speech signal acts as the most basic level of segmentation. At the same time, its local essence lies in the analysis of only adjacent vectorized frames without taking into account their higher-level context, which will be taken into account at subsequent hierarchical levels of segmentation.

At the first segmentation stage, we must exclusively receive local segments without their global context-combinatorial analysis. Due to the exceptional local properties of this segmentation, the highest speed of this algorithm is achieved in comparison with other higher-level segmentation algorithms, including those based on cluster analysis. In fact, at this stage, we carry out only the local aggregation of parametrically defined frames into low-level segments, consisting of small groups of adjacent segments. It evaluates the maximum local similarity of spectral-acoustic characteristics and forms localized dense clusters of vectors in their time sequence, interspersed with discontinuities in their density distribution.

In the process of synthesis and experimental verification of various algorithms, the most optimal results of acoustic segmentation were obtained using the developed algorithm for detecting changes in the time that is the method of distortion accumulation (Algorithm 1).

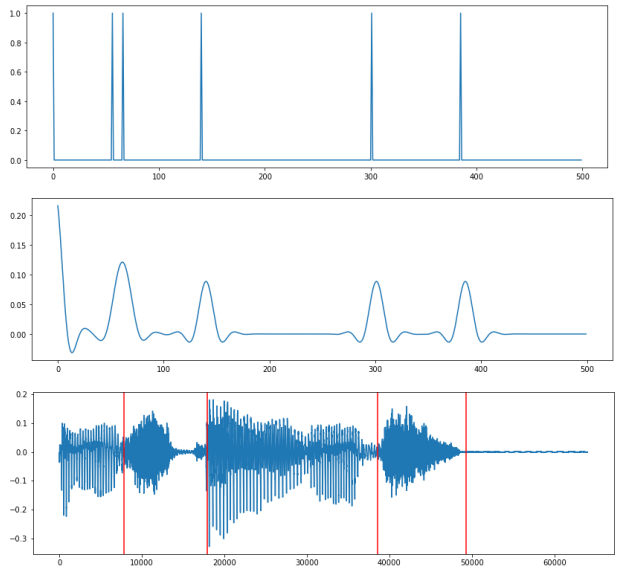

FIGURE 1 An example of the result of automatic segmentation of a speech signal using the algorithm for detecting changes in the time series of feature vectors using the error accumulation method

This algorithm takes $X$ as input data - the time series being analyzed, consisting of spectre feature vectors, as well as two parameters: $p$ - the maximum allowable deviation threshold in the formed segment, after exceeding which the local segment boundary is set and mode - sets the method for determining deviations in the segment. When *mode=Soft*, the algorithm operates in the soft mode: the next right object is added to the segment until the distance from the newly added object to the centroid of the cluster consisting of objects already added to the segment is less than the specified threshold value $p$. When *mode=Hard*, the algorithm operates in strict mode; adding the next right object to the segment is carried out as long as the distance from each already included object to the new centroid of the cluster consisting of objects already added to the segment (including the candidate object) is less than the threshold value $p$.

The proposed algorithm returns the following data as output parameters: $k$ is the number of received segments; A is a sequence of $k$ triplets $(a, b, c)$ that specifies whether each element of the time sequence belongs to a certain segment; $c$ is the centroids of the corresponding segment that are needed in subsequent calculations in case of a change in the composition of segments at subsequent levels of segmentation.

The result obtained in the form of identified boundaries is shown in FIGURE 1. An additional FIR filter was used to receive this result, allowing additional smoothing of the segment boundaries.

This filtering procedure allows one to select more general characteristics of transformations in the input sequence. For this, a software implementation of the *filtfilt* functions from the *scipy.signa* signal processing library has been utilized. FIR filtering, which performs smoothing of the obtained function of changes in the characteristics of the speech signal, uses a windowed filter. This filter function applies the linear digital filter twice: once forward and once backward. The combined filter has zero phases and twice the filtering order of the original signal. A characteristic feature of this method is the ability to regard the frequency features at the beginning and end of the processed signal. The filter parameters are adjusted for amplitude time sequences corresponding to the speech signal. The filtering procedure is initialized with the coefficients of the FIR filter (Finite Impulse Response). The *filtfilt* algorithm is used, which minimizes the initial and final transients. The coefficients for the band frequencies are selected manually. The lower and upper coefficients 0.0001, 0.2 were chosen, which correspond to the nature of the speech signal. The width of the filter window is set proportionally to the audio sampling frequency *Frequency_Rate*/10.

Algorithm 1:

function AggregationByAccumulation ($X$, $p$, *mode*)

**Input** : Raw spectral features of speech signals $X = \{x_1, \dots, x_n\}$
Sensitivity border parameter $p > 0$
Method of segmentation $mode \in \{S, R\}$

**Output** : Number of resulted segments $k$
List of left/right segmentation borders $a_i, b_i$ and segmentation centers $c_i$;
$A = \{ (a_i, b_i, c_i) \mid i = 1, \dots, k\}$
$A \leftarrow \emptyset$ ; $a \leftarrow 1$; $c \leftarrow x_1$
**foreach** $x_i \in X$
  $condition = \frac{\|x_i - c\|}{\sqrt{n}} < p$
  **if** *mode* = $Soft$ **then**
    *condition* ← *condition* &
    $\min_{x_k \in \{x_a, \dots, x_i\}} \frac{\|x_k - c\|}{\sqrt{n}} < p$
  **if not** *condition* **then**
    $b \leftarrow i - 1$; add $(a, b, c)$ into $A$; $a \leftarrow i$
    **else**
      $c \leftarrow$ **centroid** $\{ x_a, \dots , x_i\}$
      **if** $x_a \neq x_n$
        $c \leftarrow$ **centroid** $\{ x_a, \dots , x_n\}$; add $(a, n, c)$ into A

**return** : $k, A$

For the resulting smoothed function, it is analyzed for local maxima. By using the *find_peaks* function from *scipy.signal*, peaks are selected that correspond to the most significant changes in the characteristics of the speech signal. These peaks are markers for segmenting the speech signal into low-level speech analysis units.

## 2.2 MICROWAVE SEGMENTATION OF A SPEECH SIGNAL INTO SEGMENTS COMPARABLE TO THE PITCH PERIOD

New algorithm for segmenting a speech signal into segments comparable to the pitch period has been developed. The purpose of developing this algorithm is to obtain low-level atomic segments for the purpose of their subsequent hierarchical aggregation into high-level segments of a larger dimension comparable to the subphonemic or phonemic components of the speech signal.

Initially, the analyzed speech signal is uniformly segmented into mutually overlapping microwave sections comparable in dimension to the period of the fundamental tone frequency, i.e., reasonably short segments. A sufficiently small step is chosen when forming such sections, which should provide a sufficiently large degree of their mutual overlap. Then the parameterization of each of these microwave sections is carried out. Parameterization can be carried out in various ways [2, 5]. Based on the results of the parametrization, each microwave region is associated with a feature vector corresponding to it. An ordered set of such vectors form a time series.

In the next stage, the identification of local atomic clusters comparable with the fundamental tone period in the structure of the analyzed time series is carried out. The identification of such atomic local clusters is carried out by detecting discontinuities between the Euclidean distances of adjacent vectors (Algorithm 2).

Algorithm 2:

function AggregationByAttraction ($X$)

**Input** : Raw spectral features of speech signals $X = \{x_1, \dots, x_n\}$

**Output** : Number of resulted segments $k$
List of left/right segmentation borders $a_i, b_i$ and segmentation centers $c_i$;
$A = \{ (a_i, b_i, c_i) \mid i = 1, \dots, k\}$

$A \leftarrow \emptyset$ ; $a \leftarrow 1$; $c \leftarrow x_1$

**foreach** $(x_i, x_{i+1}, x_{i+2}, x_{i+3})$, $i \in \{1, \dots, n-3\}$
  $d_1 = \|x_{i+1} - x_i\|$; $d_2 = \|x_{i+2} - x_{i+1}\|$;
  $d_3 = \|x_{i+3} - x_{i+2}\|$;
  *condition* = $d_1 < d_2$ & $d_3 < d_2$
  **if** *condition* **then**
    $b \leftarrow i - 1$; add $(a, b, c)$ into $A$; $a \leftarrow i$
  **else**
    $c \leftarrow$ **centroid {** $x_a, \dots, x_i$**}**
    **if** $x_a \neq x_n$
  $c \leftarrow$ **centroid {** $x_a, \dots, x_n$**}**; add $(a, n, c)$ into A

**return** : $k, A$

The meaning of these events is that noticeable changes in the time sequence of the norms of the difference of successive vector representations mean noticeable changes in the microwave audio sequence. Thus, the sequence of the speech signal is divided into intervals for which this event is not performed, and some homogeneity/similarity of speech characteristics within the interval can be assumed. Vector representations within the obtained intervals are local clusters, from which the most characteristic representatives can be distinguished. Getting this partition is a clustering process, and typical representatives are centroids. Thus, for the obtained local clusters, centroids are calculated, the sequence of which forms the time series of the next level, and the aggregation process is repeated recursively. In practice, it is advisable to use 2 or 3 recursive repetitions to obtain segmentation with a high level of detail in the phonetic features of speech. The calculated last level of partitioning is used for comparative evaluation of this segmentation algorithm with all other proposed algorithms.

An example of the result of a microwave segmentation of a speech signal into segments comparable to the period of the fundamental tone of the analyzed speech signal is shown in FIGURE 2. The lower part of FIGURE 2 shows the segmented speech signal, and the upper part shows the sequence of received segment boundaries. The third one represents speech segments of the lowest level. Their local aggregation will be carried out at subsequent stages to obtain larger subphonemic or phonemic segments.

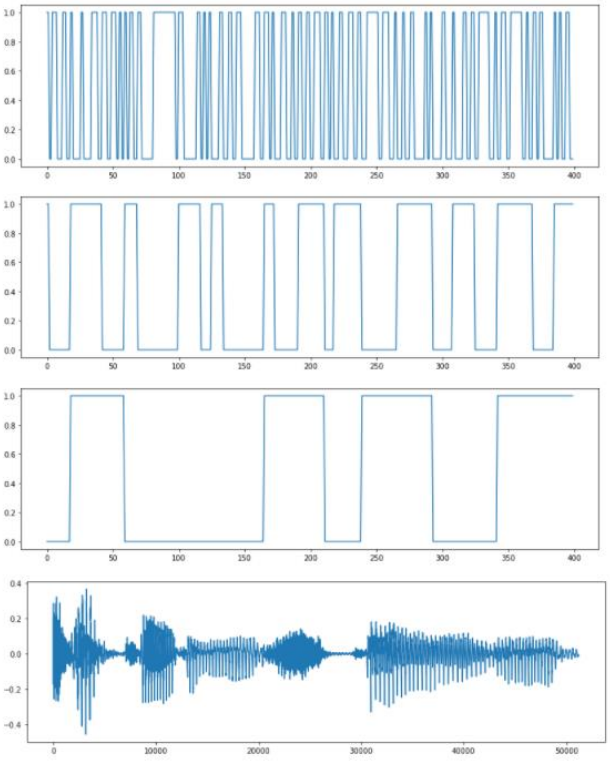

FIGURE 2 An example of obtaining a generalized segmentation result using the developed algorithm by analyzing the joint distribution of segment obtained by several different algorithms

## 2.3 LOCAL AGGREGATION OF WAVE SEGMENTS INTO LARGER SUBPHONEMIC SEGMENTS

An algorithm for multilevel sequential local aggregation of microwave segments into larger subphonemic segments has been developed. Sequential recursive local aggregation of objects in the time series is also carried out according to the principle of detecting gaps between their local clusters, but this procedure is repeated several times recursively, and

after each repetition, the centroids for all received segments are calculated, and these centroids already act as segmentation objects for the next level. For the local clusters obtained at the current level, centroids are calculated, the sequence of which forms the time series of the next level, and the aggregation process can be recursively repeated several times. Thus, the result is a multilevel segmentation of the speech signal. When moving from a lower segmentation level to a higher one, the boundaries of intermediate segments are removed. Accordingly, those boundaries that remain when moving to a higher level of segmentation are clearer, i.e., have a greater degree of unambiguity than the inter-segment boundaries that have undergone reduction. Accordingly, the lifetime of segment boundaries during multilevel aggregation and subsequent segmentation determines the degree of their importance in terms of making decisions about true phonemic boundaries.

When this procedure is used for the first time, the initial vector representations of short-term overlapping fragments of the speech signal are used as an input signal. After the distribution of these vectors over local segments is obtained at the output of this function, its centroid is calculated for each segment. An ordered set of such centroids form a new time series, which in turn is also segmented using the same procedure. Thus, segments of the next level of the hierarchy are obtained, for which centroids are also calculated, and the process can be repeated again. If this procedure is repeated many times; as a result, all segments will merge into one single segment. It should be noted that the segment boundaries of the next level are a subset of the segment boundaries of the previous level.

At this stage, the resulting total segment vector is smoothed using a FIR filter. The FIR filter parameters are adjusted so that the resulting smoothed signal peaks are located at the required distance from each other. At the final stage, the procedure for searching for peak vertices is performed and generalized segment boundaries are set in their place.

## Conclusion

This work focuses on developing highly efficient unsupervised speech signal recognition algorithms in an unknown language, which have high robustness to variability and changes in the dynamic characteristics of speech.

The main feature of the developed algorithms is their focus on the process of unsupervised machine learning. These algorithms are used to overcome the shortcomings inherent in widely used methods of supervised speech recognition by combining them. The expected results are new algorithms for solving the tasks and their software implementation as part of an automatic continuous speech recognition system.

## Acknowledgments

This research is conducted within the framework of the grant num. AP088560349 "Self-learning robust unsupervised automatic speech signal recognition".

## References

[1] Baker J et al. Updated MINDS report on speech recognition and understanding 2009 *IEEE Signal Processing Magazine* **26**(4) 78-85

[2] Noda K, Yamaguchi Y, Nakadai K, Okuno H, Ogata T Audio-visual speech recognition using deep learning 2015 *Applied Intelligence* **42**(4) 722-737

[3] Karpov A An automatic multimodal speech recognition system with audio and video information 2014 *Automation and Remote Control* **75**(12) 2190-2200

[4] Solera-Urena R, Garcia-Moral A, Pelaez-Moreno C, Martinez-Ramon M, Diaz-de-Maria F Real-Time Robust Automatic Speech Recognition Using Compact Support Vector Machines 2012 *IEEE Transactions on Audio Speech and Language Processing* **20**(4) 1347-1361

[5] Baker J, Li D, Glass J., Khudanpur S, Chin-hui L, Morgan N, O'Shaughnessy D Developments and directions in speech recognition and understanding 2009 *IEEE Signal Processing Magazine* 26(3) 75-80

[6] Champion C, Houghton S Application of continuous state Hidden Markov Models to a classical problem in speech recognition 2016 *Computer Speech and Language* **36** 347-364

[7] LeCun Y, Bengio Y, Hinton G. Deep learning 2015 *Nature* **521**(7553) 436-444