УДК 004.934
МРНТИ 16.31.21 DOI 10.55452/1998-6688-2021-18-3-89-94

A REVIEW: METHODS OF AUTOMATIC SPEECH SEGMENTATION

PAK A.A.[1], ZHUMAGELDIKYZY A.[2], ERMEKOVA N.S.[3]
[1]Institute of Information and Computational technologies, 050000, Almaty, Kazakhstan
[2]Kazakh-British technical university, 050000, Almaty, Kazakhstan
[3]Zhetysu university named after I.Zhansugurov, 040000, Taldykorgan, Kazakhstan

Abstract. *Segmentation is a process of dividing a speech signal into the basic units of language. Segmentation of the speech signals is one of the most important tasks in automatic speech processing systems. This paper proposes a review of methods of automatic speech segmentation. Moreover, methods of wavelet and Hilbert-Huang transformations and techniques based on hidden Markov models are considered.*

Keywords: *Speech signals; Speech segmentation; Automatic segmentation methods; Method of discrete wavelet transform; Hilbert-Huang transform; hidden Markov models.*

ШОЛУ: СӨЙЛЕУДІ АВТОМАТТЫ СЕГМЕНТТЕУ ӘДІСТЕРІ

ПАК А.А.[1], ЖҰМАГЕЛДІҚЫЗЫ А.[2], ЕРМЕКОВА Н.С.[3]
[1]Ақпараттық және Есептеуіш технологиялар институты, 050000, Алматы, Қазақстан
[2]Қазақстан-Британ техникалық университеті, 050000, Алматы, Қазақстан
[3] И.Жансүгіров атындағы Жетісу университеті, 040000, Талдықорған, Қазақстан

Аңдатпа. *Сегменттеу – сөйлеу сигналын тілдің негізгі бірліктеріне бөлу процесі болып табылады. Сөйлеу сигналдарын сегментациялау автоматты түрде сөйлеуді өңдеу жүйелеріндегі маңызды міндеттердің бірі болып саналады. Бұл мақалада сөйлеуді автоматты түрде сегментациялау әдістеріне шолу жасалады. Сонымен қатар вейвлет және Гильберт-Хуанг түрлендіру әдістері мен жасырын Марков модельдеріне негізделген техникалары қарастырылады.*

Түйінді сөздер: *сөйлеу сигналдары, сөйлеуді сегментациялау, автоматты сегментациялау әдістері, дискретті вейвлет түрлендіру әдісі, Гильберт-Хуанг түрлендіруі, жасырын Марков модельдері.*

ОБЗОР: МЕТОДЫ АВТОМАТИЧЕСКОЙ РЕЧЕВОЙ СЕГМЕНТАЦИИ

ПАК А.А.[1], ЖУМАГЕЛДЫКЫЗЫ А.[2], ЕРМЕКОВА Н.С.[3]
[1]Институт информационных и вычислительных технологий, 050000, Алматы, Казахстан
[2]Казахстанско-Британский технический университет, 050000, Алматы, Казахстан
[3]Жетысуский университет им. И.Жансугурова, 040000, Талдыкорган, Казахстан

Аннотация. *Сегментация – это процесс разделения речевого сигнала на основные языковые единицы. Сегментация речевых сигналов – одна из важнейших задач в системах автоматической обработки речи. В данной статье предлагается обзор методов автоматической сегментации речи. Кроме того, рассматриваются методы преобразований вейвлетов и Гильберта-Хуанга, а также техники, основанные на скрытых Марковских моделях.*

Ключевые слова: *речевые сигналы, сегментация речи, методы автоматической сегментации, метод дискретного вейвлет-преобразования, преобразование Гильберта-Хуанга, скрытые Марковские модели.*

Introduction

Research in the field of speech signal processing is an active progress. Despite the high speed of computer technology and information technology development, the main problems of speech applications are still relevant. One of the most important tasks in automatic speech processing systems is the task of segmentation by the phonetic transcription of the language[7]. The main reason for the segmentation research is the complexity of the structure of the speech signal: a huge variety of phonetic units of the language, intonation colors, personal characteristics of the speaker is aggravated by a variety of external factors that affect the recording and transmission of voice. As a result, speech signals are rather difficult to investigate and describe in detail using mathematical models. Digital processing methods imply the possibility of their use for solving problems of speech signal processing.

Processing and transmitting speech signals is an important component of modern radio engineering and several related areas, such as computer science, cybersecurity, etc. The proportion of data transmitted in the form of speech signals remains significant, and most of them are digital. Besides, speech and sound signals are important components of video signals. The role of speech signals used for personal identification in biometric systems is also significant. In conclusion, there is a huge relevance of research and knowledge in the field of digital processing of speech signals. [1]

Speech processing is a field of science that deals with filtering, amplification, and extraction of information, coding, compression, and restoration of speech. Processing in speech recognition systems includes the following tasks:

- filtering and noise suppression;
- segmentation into informative areas;
- determination of informative parameters;
- recognition. [2]

This article provides an overview of current research in the speech segmentation task area. Sentence segmentation has great importance for speech understanding applications—from parsing and information extraction at the more basic level to machine translation, summarization, and question answering at the application level. Sentence boundaries are also important for aiding the human readability of the output of automatic speech recognition systems. [3]

About speech signals

Before embarking on a detailed discussion of speech segmentation and its methods, the speech signal, the source of which is the sounds generated by the human articulatory apparatus must be taken into account. The system can be thought of as a tube with a variety of autonomously moving barriers inside it, such as the tongue, lips, and vocal cords, that can change its cross-section. When pressured air passes through this tube, it produces sound. The air pressure may be controlled extremely precisely in this scenario, and the specified barriers can be shifted extremely quickly. As a result, well-known sounds arise clicks, whistles, and others that a person is capable of. In any language, about 40 or 50 types of sounds related to speech, the so-called phonemes, can be distinguished. [4] Phonemes are the basic units of the language in speech and in computational linguistics, where there are different linguistic units such as morphemes, lexemes, etc., which form a complex hierarchy of interactions. In this regard, the methods of computer analysis should be able to take into account a variety of phonemes.

A particular speech sound is created by a specific pattern of muscle movements in the vocal channel. Vowels and consonants are the two basic categories of speech sounds that can be identified.

In the speech, in addition to sequences of sounds, there is also a pause, the presence of which can indicate the end of an utterance or thought. Pause is usually accompanied by silence or noise [5].

Segmentation and its methods

After the general idea of speech sound construction, the segmentation can be begun. Due to the peculiarities of the human brain in the field of processing verbal information that is received through speech, the utterance end of the boundaries could be distinguished. In Natural Languages, speech is the sequential link of phonemes [6]. Speech segmentation can be defined as the process of finding the limits (with specific characteristics) in natural spoken language between words, syllables, or phonemes [5].

Moreover, there are two main methods used for the segmentation of speech signals: *manual* segmentation and *automatic* segmentation. Manual segmentation can be used in research systems and at the pre-development stage. However, it requires a significant investment of time and effort: firstly, there are no pauses between words in continuous speech, and secondly, coarticulation, which also occurs at the border of sequentially produced sounds, which greatly facilitates the correct perception and understanding of speech, but makes it difficult to find the boundaries of segments. In addition, it is almost impossible to accurately reproduce the results of manual segmentation due to the subjectivity of human auditory and visual perception. [7] Consequently, automatic segmentation is used for the segmentation of speech signals.

Automatic segmentation

As it has been said before, segmentation of speech can be done into basic units like phonemes, words, or syllables. The automatic segmentation of speech using only the phoneme sequence is an important task, especially if manually pre-segmented sentences are not available for training [6].

Each task of processing speech signals can only be realized using certain methods. Depending on the area of processing, the methods should be divided into three areas: frequency, time, and frequency-time, where segmentation refers to the frequency-time area.

Time-domain processing methods consist of determining characteristic points of a speech signal and then using them for analysis. The main disadvantage of time-domain processing methods is the ambiguity in the extraction of key points caused by noise and offsets.

Frequency domain processing techniques are based on the use of all data samples recorded in the speech signal. The use of methods in the frequency domain facilitates the process of speech signals with sufficiently high accuracy. The disadvantages of processing in the frequency domain include low adaptability to the local properties of signals, insufficiently high spectral resolution, and relatively high computational costs.

Time-frequency domain processing techniques incorporate all the advantages of time and frequency analyzes with minimal manifestations of their disadvantages. [2] There are four methods in time-frequency domain processing: Fourier transforms, wavelet transforms, linear prediction analysis, and Hilbert-Huang transform.

Compare with wavelet transform and Hilbert-Huang transform, the Fourier transform(FT) and parameterization with linear prediction coefficients are not suitable for analyzing non-stationary signals, due to the loss of temporal feature information. [10] For instance, the FT reveals global information about the signal's frequencies but does not give a notion of the signal's local features when its spectral composition changes rapidly over time. Moreover, The FT cannot analyze the frequency characteristics of a signal at arbitrary times. These shortcomings stimulated the development of the wavelet transforms. [15]

Analysis via discrete wavelet transform method

Neurophysiological studies of the human brain claim that an important step in the pipeline of speech recognition is frequency analysis (Daubechies 1992). It's reasonable to imitate the step at machine speech recognition, discrete wavelet transform(DWT) is a good candidate method due to its universality in digital signal processing. The DWT is a special case of the Wavelet Transform (WT) that gives a concise time and frequency representation of a signal that may be computed quickly [11]. The accuracy 16 bit is well enough to build the wavelet spectrum of speech. Besides WT there is discrete Fourier transform(FT) in a family of frequency transformation methods. The crucial differences between DWT and DFT are the ability of DWT to localize time intervals with specific frequency patterns, in other words, at what scale the pattern occurred at an original signal, that solves the task of search speech parameters which are important for the human hearing system (Wang and Narayanan 2005). The model of DWT can be presented as

$$s(t) = \sum_i c_{m+1,i} \Phi_{m+1,i}(t) ,$$

where $\Phi_{m+1,i}(t)$ is a *i*-th wavelet function at *(m+1)*-th scale level; *s(t)* is resulting signal function.

Additionally, there are coefficients of the lower level:

$$c_{m,n} = \sum_i h_{i-2n} c_{m+1,i}$$

$$d_{m,n} = \sum_i g_{i-2n} c_{m+1,i} \quad ,$$

where h and g are the constants that depend on the pair of scale function Φ and wavelet Ψ. In such a manner, the mentioned equations decompose the original signal by filtering it against the base wavelet function Ψ. In the pipeline of DWT constant coefficients are collected into a vector $(d_m, d_{m-1,} d_{m-2,} \dots, d_{1,} c_1)$. The coefficient of other scales calculated recursively according to the mentioned above iterative equations. As a result, DWT includes the hierarchical step of frequency pattern analysis that leads to a multi-resolution analysis of the original signal. The advantage of DWT is a fast computational scheme. The banking filter helps to generate the wavelet spectrum, which has a tree-like structure. In other words, there is the sequence of cascading filtering and downsampling operations. The root of the tree is wavelet coefficients of the original signal, downstream levels of the tree are wavelet coefficients after downsampling.

Johnson Ihyeh Agbinya presented a voice compression approach based on wavelet techniques. Speech compression includes both voiced and unvoiced speech, as well as a variety of wavelet types. Wavelets are used in this procedure, and low-frequency coefficients are used. Energy in bands is used to detect the voiced and unvoiced parts of a speech signal. [12]

By breaking down the wavelet, S. Ratsameewichai, N. Theera-Umpon, J. Vilasdechanon, S. Uatrongjit, and K. Likit-Anurucks were able to separate the speech into low and high-frequency components. They then used the energy contour to determine the phoneme's limits. They experimented with 1,000 syllables of data collected from ten speakers. The accuracy rates are 96.0, 89.9, 92.7, and 98.9% for initial consonant, vowel, final consonant, and silence, respectively. [13]

Bartosz Zioko described the Wavelet technique in their publication, which is used to detect phonemes based on power variations. Using spectral analysis of speech, the information from the speech signals may be efficiently retrieved. The DWT can be used to perform spectral analysis. To determine the beginning and end of phonemes, the power is evaluated in several frequency sub-bands. Power transitions in wavelets can be used to detect the boundaries of phonemes. [14]

Analysis via Hilbert-Huang Transform

The Hilbert – Huang transform (HHT) represents the decomposition of a signal into empirical modes(EM), followed by the application of the derived components of the Hilbert transform expansion to get integral information on the signal's amplitude-frequency-time parameters. The Empirical Mode Decomposition (EMD) method is intended for the analysis of non-stationary and non-linear processes. Compare with Fourier and wavelet analysis, EMD is direct, intuitive, adaptive with a posteriori determined basis that depends on the signal data and is built using the decomposition method. [15] The main advantage of this method is its high adaptability, which is manifested in the fact that the basic functions of sound decomposition are extracted directly from the original signal itself and allow only its inherent features to be taken into account [2].

HHT includes two main stages:

1. Decomposition of a signal into components [16]:

$$s(t) = \sum_{i=1}^{I-1} imf_i(t) + r_I(t) \quad ,$$

where $imf_i(t)$-empirical modes(EM); $r_I(t)$-decomposition residual, $i = 1,2,\dots,I$-number of EM.

2. Formation of the obtained empirical modes of the Hilbert spectrum [17]:

$$HHT(t) = \sum_{i=1}^{T} a_i^{\,2}(t) \cdot e^{q\int \omega_k(t)dt} \quad ,$$

where $a_i(t) = \sqrt{imf_i(t)^2 + IMF_i(t)^2}$-modulus of the instantaneous value of the signal amplitude of each EM; $imf_i(t)$-EM of signal; $IMF_i(t) = \frac{1}{\pi}\int \frac{imf_i(\tau)}{t-\tau} d\tau$-Hilbert-coupled EM signal; τ-time shift proportional to the phase of the signal; $\omega(t) = 2\pi f j$-cyclic frequencies of each EM; j-imaginary unit.

The values $a(t)$ and $\omega(t)$ are determined from the analytical signal of $Z_i(t) = imf_i(t) + jIMF_i(t)$ each EM.

The speech signal is represented in the frequency-energy-time domain as a consequence of HHT, which allows for the discovery of hidden modulations and areas of energy concentration, as well as the analysis of both global and local aspects of signals at lower computational costs.

Analysis via Hidden Markov Model-based techniques

Hidden Markov Models (HMM) are widely used in speech recognition tasks due to their high performance in recognition and relatively small computational complexity in the field of speech recognition. Nevertheless, techniques based on HMM are more practiced in the field of segmentation.

J. Dines, S. Sridharan, and M. Moody have discussed the features of their automatic speech segmentation system, which are used in their speech synthesis study. It was built using training procedures tuned for the segmentation job and a Hidden Markov Model phone recognizer, which identified the differences in voice segmentation estimation methodologies. The capacity of their technology to produce high-reliability speech segmentation is demonstrated through system evaluation. [18]

Techniques for improving the accuracy of automatic phonetic segmentation based on HMM acoustic-phonetic models were presented by A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liberman. It was found that applying more powerful statistical models for border correction is more conditioned on phonetic context and duration variables enhanced test results. Moreover, the discovery of merging various acoustic front-ends resulted in an extra gain inaccuracy, and that conditioning the combiner on phonetic context and side information improves outcomes, which reduced segmentation errors on the TIMIT corpus by nearly half, from 93.9 percent to 96.8 percent boundary correctness with a 20-ms tolerance. [19]

Conclusion

In this paper, methods for speech segmentation and hidden Markov model-based techniques are described. Among the investigated methods, two are chosen to be analyzed - the Hilbert-Huang transform and wavelet transform. The Fourier transform and parameterization with linear prediction coefficients are not suitable for non-stationary signals analysis. The Hilbert-Huang is less used compared with wavelet transform due to the number of its applications for solving practical problems. The most accurate and efficient methodology researched was the hidden Markov model. Consequently, the Markov model has to be observed wider for the speech segmentation field.

In addition, researches related to the segmentation of Kazakh speech have not been considered due to the less use. Nevertheless, the researches related to Russian speech have been analyzed recently. The methods discussed above can be applied for the segmentation of Kazakh speech.

REFERENCES

1. A. I. Topnikov. BBK 387-013ya73 T58 Rekomendovano Redakcionno-izdatel'skim sovetom universiteta v kachestve uchebnogo izdaniya. Plan 2018 goda. – 2018.
2. A. K. Alimuradov, P. P. Churakov. Obzor i klassifikaciya metodov obrabotki rechevyh signalov v sistemah raspoznavaniya rechi //Izmerenie. Monitoring. Upravlenie. Kontrol'. – 2015. – №. 2 (12).
3. D. Jones. et al. Measuring human readability of machine-generated text: three case studies in speech recognition and machine translation //Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. – IEEE, 2005. – T. 5. – C. v/1009-v/1012 Vol. 5.
4. Hant E. Iskusstvennyj intellekt. 1978 g.- 558 s
5. A. E. Sakran. et al. A Review: Automatic Speech Segmentation //International Journal of Computer Science and Mobile Computing. – 2017. – T. 6. – №. 4. – C. 308-315.

6. K. Geetha, R. Vadivel. Phoneme Segmentation of Tamil Speech Signals Using Spectral Transition Measure //Oriental Journal of Computer Science and Technology. – 2017. – Т. 10. – №. 1. – С. 114-119.
7. O. A. Vishnyakova, D. N. Lavrov. Avtomaticheskaya segmentaciya rechevogo signala na baze diskretnogo vejvlet-preobrazovaniya //Matematicheskie struktury i modelirovanie. – 2011. – №. 2 (23).
8. I. Daubechies. Ten lectures on wavelets. – Society for industrial and applied mathematics, 1992.
9. D. Wang, S. Narayanan. Piecewise linear stylization of pitch via wavelet analysis //Ninth European Conference on Speech Communication and Technology. – 2005.
10. K. K. Tomchuk. Segmentaciya rechevyh signalov dlya zadach avtomaticheskoj obrabotki rechi : dis. – S.-Peterb. gos. un-t aerokosm. priborostroeniya, 2017.
11. G. Tzanetakis, G. Essl, P. Cook. Audio analysis using the discrete wavelet transform //Proc. Conf. in Acoustics and Music Theory Applications. – 2001. – Т. 66.
12. J. I. Agbinya. Discrete wavelet transform techniques in speech processing //Proceedings of Digital Processing Applications (TENCON'96). – IEEE, 1996. – Т. 2. – С. 514-519.
13. S. Ratsameewichai. et al. Thai phoneme segmentation using dual-band energy contour // Proceedings of the IEEK Conference. – The Institute of Electronics and Information Engineers, 2002. – С. 110-112.
14. B. Ziółko. et al. Wavelet method of speech segmentation //2006 14th European Signal Processing Conference. – IEEE, 2006. – С. 1-5.
15. Yu. E. Ul'yanova, R. G. Babenko, A. V. Chernov. Chastotno-vremennye preobrazovaniya, ispol'zuemye v cifrovoj obrabotke signalov //Global'naya yadernaya bezopasnost'. – 2015. – №. 3 (16).
16. N. E. Huang. et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis //Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences. – 1998. – Т. 454. – №. 1971. – С. 903-995.
17. E. Huang. Huang. Hilbert-Huang Transform and its application. Interdisciplinary mathematical sciences / E. Huang. Huang, S. P. Samuel. Shen // Interdisciplinary Mathematical Sciences. Book 5. World Scientific Publishing Company. – Sep. 2005. – 324 p.
18. J. Dines, S. Sridharan, M. Moody. Automatic speech segmentation with hmm //Proceedings of the 9th Australian Conference on Speech Science and Technology. – 2002. – С. 544-549.
19. A. Stolcke. et al. Highly accurate phonetic segmentation using boundary correction models and system fusion //2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2014. – С. 5552-5556.

Information about authors:

1. Pak A.A. – Institute of Information and Computational Technologies, Institute of Information and Computational Technologies, st. Pushkin 125b, Almaty
2. Zhumageldikyzy A. – Kazakh-British Technical University, st. Tole bi 59, Almaty
3. Ermekova N.S. – Zhetysu University named after I.Zhansugurov, st. I. Zhansugurov 187a, Taldykorgan