Advances in Economics, Business and Management Research, volume 138
2nd International Scientific and Practical Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth" (MTDE 2020)

The Usage of Digital Technology in Speech Segmentation

Rzayeva Zumrud Elyar

Ganja State University, Ganja, AZ2001, Azerbaijan
Email: lingui80@mail.ru

ABSTRACT
The digital technology means applied in the segmentative analysis of speech signals is considered in the article. To create an informative base about the existing methods of segmentation using digital technologies, which is the purpose of this article, a descriptive method of linguistics is used. For the first time, the article briefly analyzes methods of segmentation of speech signals using digital technologies, which include such analysis methods as spectral, cep spectral analysis and wavelet processing of speech signals, and the main characteristics of each method are indicated in terms of comparison. Summarizing the information presented during the study, the conclusion indicates that the spectral analysis of speech signals is acceptable for forensic studies, and the cepstral analysis of speech signals helps to determine the psycho-emotional state of the speaker. The specific software for segmenting speech signals is considered in a laboratory environment.
Keywords: *segmentation, speech signals, digital technologies, spectral analysis, segmentation methods*

1. INTRODUCTION

The importance of quantity and quantitative methods is especially appreciated in modern society of high technology and scientific and technological progress. It is impossible to imagine the full, adequate life of a modern person without his understanding of the basic principles of the quantitative structure of the world. At the same time, without improving knowledge of quantity, its laws and functions, the further development of life is impossible.

The quality differentiation had led to the need of its accurate measurement and calculation. This contributed to the emergence of a numerical account - the most objective characteristic of reality. By using numbers, a man formulated the laws of nature, explained them; this became a prerequisite for the emergence of science. With the development and complication of scientific knowledge, scientists began not only to use quantity and counting for explaining of the phenomena of surrounding reality, but also to study the quantity, identify its laws, and find the numerical characteristics of objects of the material world. Quantitative methods began to be applied not only in mathematical sciences, but also in the humanitarian sciences.

In various studies in speech processing (speech recognition, speech compression, speaker identification by voice, etc.), one has to deal with the problem of segmentation of speech units at different language levels. Moreover, you can find many materials about the types and characteristics of speech segmentation in the scientific literature. The appropriateness of the selected type of segmentation is determined by several factors. This can include a concrete speech-processing task to solve the selected specific task, study time, requirements and level of accuracy, etc. Based on this, specialists from related fields have developed a number of methods and software for conducting such experimental studies. In this article, we are going to consider all the various ways and possibilities that modern digital technologies can provide us with, which is the scientific novelty of this study. The relevance of the study lies in the fact that this topic is new and in demand for speech theory. We believe that the materials of this study can be useful for researchers in this field and acquaint them with the latest achievements in linguistics field.

The need for segmentation of voice messages arose along with the need to detect exact boundaries separating speech periods and noise periods, which is very important when transmitting voice information over communication channels.

A special role in speech technology is enacted by the area related to the automatic recognition and perception of human speech. Active research in the field of speech recognition had begun about 60 years ago. The organizations such as Bell Laboratories, RCA Labs, University College in England, MIT Lincoln Labs, Research Institute for Long-distance Communications (Leningrad), Institute for Information Transmission Problems of the Russian Academy of Sciences carried out the works on this field.

2. BACKGROUND

The development of digital processing methods has expanded the capabilities of speech processing. We can see it in the works of famous foreign authors, whose works

have become "classics" in the fields of speech processing and digital signal processing. These authors are Rabiner L. R. and Gold B. [Rabiner, Gold 1978], Rabiner L. R. and Shafer R. V. [Rabiner, Shafer 1981], Oppenheim A. [The use of digital signal processing 1980], Markel J. D. and Gray A. Kh. [Markel, Gray 1980].
Segmentation can be context-sensitive and context-independent. According to the principle for determining the boundaries of segments, there is segmentation into fragments of a fixed duration and phoneme segmentation. The most difficult task is context-independent phoneme segmentation. In order to understand the essence of the application of digital technology in the segmentation of speech units, it is necessary to consider methods of segmentation of speech units and signals. The following methods of segmentation of speech signals are marked in the scientific literature: spectral analysis of the speech signal, cepstral analysis of the speech signal, usage wavelet transforms in the processing of speech signals, correlation analysis of the speech signal.
Let us look through all of this types.

2.1. Spectral analysis of the speech signal

In general case, the signal incurred to a discrete Fourier transform in the spectral analysis of speech signals, then for the resulting spectrum a logarithmic scale is performed in the spaces of amplitudes and frequencies (chalk conversion frequency), the spectrum is smoothed and its circumflex is extracted. This allows taking into account the decrease in the information content of the high-frequency components of the signal, as well as the logarithmic sensitivity of human hearing.
It is necessary to emphasize the negative impact of the variability of speech utterance, speaking about the methods used in the spectral region. Under the influence of various factors such as the emotional mood, physical condition, a person produces SS for the same phrases or words, significantly differing in spectral-temporal characteristics. Besides it, the interpenetration of neighboring sounds is changed in consequence of changing the tempo and volume of the pronunciation [Krasheninnikova 2011: 188–191].
The method of phonetic spectral analysis is successfully used in forensics. It is considered that only “persistent signs can be used to establish the identity presence or absence of a person for identification” [Kaganov 2013: 165]. “The methodology for obtaining and evaluating the values of the formant characteristics of vowels can be effective in these cases, which allows us to compare phonetic contexts that are traditionally defined as mismatched, but those in which the acoustic quality of the sounds under study is preserved” [Kaganov 2013: 166].
In view of the foregoing, it is necessary to use segmentation algorithms synchronized with the structure of the speech signal, while the sustainability of algorithms to the variability in speech utterance must be provided.

2.2 Cepstral analysis of the speech signal

Most modern automatic speech recognition systems are based on extracting the characteristics of the current state of the human speech tract, rather than the excitation signal, since the parameters of the speech signals obtained in the first case perform a distinction function better. For separation the excitation signal from the signal of the vocal tract, cepstral analysis is betaken [Shariy 2008: 536-541].
A cepstrum is the inverse Fourier transform of the logarithm of a power spectrum. The logarithm operation leads to separation in the spectral region of terms containing information about the probing **signal** and the delay time of one echo **signal** relative to another. In the field of speech signal processing, cepstral analysis has gained wide practical popularity due to the explanation of compressing information about the signal when transmitting from the temporal to the frequency sphere of processing [Huang 2001]. Cepstral analysis is based on the highlighting of cepstral coefficients on the chalk scale, called small-frequency cepstral coefficients. They include two basic concepts: cepstrum and chalk scale. A cepstrum is a discrete-cosine transformation of the amplitude spectrum of a signal on a logarithmic scale [Alimuradov 2018: 91].
The use of cepstral analysis in the classical method can improve the efficiency of disclosure and assessment of the psycho-emotional state. The best results are achieved by subtracting the noise background from the initial speech signal and this method can be successfully tested in remote monitoring systems for assessing the emotional state.

2.3 The usage of wavelet transforms in the processing of speech signals

The Fourier transform and parameterization by linear prediction coefficients are not suitable for the analysis of non-stationary signals, since in this case the signal information about the temporal features is lost [Petrov 2008: 135–136].
Wavelet signal processing provides the possibility of the efficient compression of signals, including speech, and their recovery with low loss of information.
In speech signals, there are both phoneme fragments with a relatively slow change in the spectral representation, and sections of the fast reconstruction of the speech apparatus (interfonemic transitions, explosive phonemes) and, accordingly, the rapid changes in the signal spectrum. Such instability fragments make it reasonable to use wavelet analysis to study the properties of a speech signal. The wavelet spectrograms obtained in the consequence of the wavelet transform contain information about the formant frequencies and the harmonic structure of the initial speech signal [Basile 1993: 169–174]. The basic functions of the wavelet dissociation have the ability to detect both frequency and temporal characteristics, which as a result makes it possible to isolate and localize the

temporal features of speech signals [Petrov 2008: 135–136].
The transition from one phoneme to another is caused by a change in the configuration of the speech apparatus, which is reflected in a harsh change in the wavelet coefficients at one or several scales of dissociation of the SS [Yermolenko 2003: 306–310].
Consider the 3 most popular and current software packages that are appropriate to use in the analysis and processing of speech signals, compression, noise removal or reconstruction in particular: 1) System Expansion Pack of MATLAB 6.0/6.1/6.5 Wavelet Toolbox 2/2.1/2.2; 2) Mathcad software package; 3) Wavelet Explorer system Mathematica. However, despite the fact that each of these software systems has wide functionality for working with speech signals, we can say that they have distinctive features, and have visible differences in instrumentation and visual design.

2.4 Correlation analysis of the speech signal

The correlation analysis of signals is widely used in the segmentation of a speech signal, as it allows estimating the magnitude of the energy of the analyzed fragment, the fact of the presence of vocalization, its frequency localization [Zhuikov 2010: 83-89]. The main maximum of the autocorrelation function (ACF) corresponds to the energy of the analyzed fragment of the speech signal; its decrease rate characterizes the presence of the noise component in the speech signal. For the voiced fragments of the autocorrelation function, it has a characteristic form similar to that of the autocorrelation function of a rectangular radio pulse.
By the value of the first local maximum of the autocorrelation function, one can judge the presence of vocalization and the average value of the duration of the OT period in the speech fragment under consideration [Rabiner, Shafer 1981: 82]. Using the autocorrelation method, the fundamental frequency was extracted when creating the phonetic database of the Institute of the Russian Language of the Russian Academy of Sciences [Kodzasov 1996].
By the meaning of the first local maximum of the autocorrelation function, one can judge the presence of vocalization and the average value of the duration of the period in the considered speech fragment [Rabiner, Shafer 1981: 82]. By the autocorrelation method, the extraction of main frequency tone was carried out on creating the phonetic database of the Institute of the Russian Language of the RAS [Kodzasov 1996].
The autocorrelation function is relatively stable to noise, however, to estimate the periodicity of a voiced fragment requires a sufficiently large interval analysis. As a result, when considering transitional sections of voiced segments, accompanied by a rapid change in the structure and frequency of vibrations of the vocal cords, local maxima become unexpressed.
In the speech analysis, an approach, defining any time-limited non-stationary signal as a particular implementation of an infinitely long stationary signal is used. However, the direct application of the result (2) to speech signals does not allow achieving the maximum efficiency of the algorithm due to the strong non-stationarity of speech signals. Separating words into phonemes as part of phonetic analysis is a separate and complex task of practical linguistics.
The simplest and most computationally efficient way of phonetic analysis of the word structure is to divide the word into non-overlapping adjacent segments of fixed length. In this case, the a priori database of the dictionary will contain not one, but several matrices for each word, calculated for the corresponding segments, and the calculation of the decisive statistics will be carried out by averaging the decisive statistics calculated for all segments.

3. CONCLUSION

Contextually independent phoneme segmentation is the most difficult task in speech segmentation.
The spectral segmentation of speech signals is more effective in forensic science, which allows the physiological characteristics of criminals or suspected persons to be determined by speech characteristics.
Using segmentative units the cepstral analysis of a speech signal helps determine the mental state of the speaker.
In wavelet processing of signals, such software as MATLAB 6.0/ 6.1/ 6.5 Wavelet Toolbox 2.2.1 / 2.2, Mathcad and Wavelet Explorer of Mathematica are used very successfully.

4. DISCUSSION

The materials and conclusions of this study may be useful for further research in this direction and in the development of new software for segmentation of speech signals in the laboratory.

REFERENCES

[1] Alimuradov A. K. et al. Assessment of the psycho-emotional state of a person based on decomposition into empirical modes and cepstral analysis of speech signals. // Bulletin of Penza State University/ — [*Alimuradov A. K. i dr. Otsenka psikhoeiotsional'nogo sostoyaniya cheloveka na osnove dekompozitsii na empiricheskie mody i kepstral'nogo analiza rechevykh signalov. // Vestnik Penzenskogo gosudarstvennogo universiteta*]. Vol. 22, No 2, 2018, PGU, Penza, pp. 89-95. (in Russian).

[2] Basile, P. The time-scale transform method as an instrument for phonetic analysis / P. Basile, F. Cutugno, P. Maturi, A. Piccialli // Visual representations of speech signals / Chicester, UK : John Wiley & Sons, 1993. – Chapter 13. – pp. 169–174

[3] Huang X. Spoken Language Processing // Guide to Algorithms and System Development. – Prentice Hall, Upper Saddle River, 2001.

[4] Kaganov A. Sh. Instrumental study of the spectral characteristics of speech in the task of forensic identification of the speaker: theoretical foundations and research technology. // Bulletin of the Russian State Humanitarian University. Series: Literary Studies. Linguistics. Cultural science. — [*Kaganov A. Sh. Instrumental'noe issledovanie spektral'nykh kharakteristik rechi v zadache kriminalisticheskoy identifikatsii govoryashchego: teoreticheskie osnovaniya i tekhnologiya issledovaniya. // Vestnik RGGU. Seriya: Literaturovedenie. Yazykoznanie. Kul'turologiya*]. 2013. № 8 (109) RGGU, Moscow. pp. 164-181. (in Russian)

[5] Kodzasov S. V. Phonetic database of the Institute of Natural Sciences of the Russian Academy of Sciences as a source of prosodic information / S. V. Kodzasov // Prosodic system of Russian speech. — [*Kodzasov S. V. Foneticheskaya baza dannykh IRYa RAN kak istochnik prosodicheskikh svedeniy // Prosodicheskiy stroy russkoy rechi*] Moscow, Institut russkogo yazyka RAN,1996. – 256 p. (in Russian)

[6] Krasheninnikova, N. A. The main factors preventing recognition of speech commands // Simbirsk Scientific Bulletin/ — [*Krasheninnikova N. A. Osnovnye faktory, meshayushchie raspoznavaniyu rechevykh komand // Simbirskiy nauchnyy vestnik*] – 2011. – Vol 3, No 1, – p. 188–191. (in Russian).

[7] Markel J. D. Linear Prediction of Speech: [trans. from English.]. — [*Markel Dzh. D. Lineynoe predskazanie rechi : (per. s angl.)*] Moscow, Svyaz',1980. – 308 p. (in Russian).

[8] Petrov A. A. Isolation of the signs of a speech signal based on wavelet analysis // Proceedings of the VI All-Russian Scientific and Practical Conference Youth and Modern Information Technologies. — [*Petrov, A. A. Vydelenie priznakov rechevogo signala na osnove veyvlet-analiza / A. A. Petrov // Sbornik trudov VI Vserossiyskoy nauchno-prakticheskoy konferentsii Molodezh' i sovremennye informatsionnye tekhnologii*]/ Thomsk: TPU, 2008. – pp. 135–136. (in Russian).

[9] Rabiner L. R. Theory and application of digital signal processing. — [*Rabiner L. R. Teoriya i primenenie tsifrovoy obrabotki signalov*]. – Moscow, Mir, 1978. – 848 p.

[10] Rabiner L. R. Digital processing of speech signals: [trans. from English.]. — [*Rabiner L. R. Tsifrovaya obrabotka rechevykh signalov: (per. s angl.)*]. – Moscow, Radio I svyaz', 1981. – 496 p. (in Russian).

[11] Shariy T.V. On the problem of parameterization of a speech signal in modern speech recognition systems. // News of Donetsk National University. — Ser. A: Natural sciences. — [*Shariy T. V. O probleme parametrizatsii rechevogo signala v sovremennykh sistemakh raspoznavaniya rechi // Visnik Donets'kogo natsional'nogo universitetu. – Ser. A: Prirodnichi nauki.*] — Vol. 2. – 2008. – pp. 536–541. (in Russian).

[12] The use of digital signal processing: [trans. from English.] / Ed. E. Oppenheim. — [*Primenenie tsifrovoy obrabotki signalov : [per. s angl.] / pod red. E. Oppengeyma*] – Moscow, Mir, 1980. – 552 p. (in Russian).

[13] Yermolenko T. V. Segmentation of a speech signal with application of fast wavelet-transformation // International Journal on Information Theories and Applications. – 2003. – Vol. 10. – No. 3. – pp. 306–310

[14] Zhuikov V. Ya. Algorithm for automatic classification of speech segments based on autocorrelation and energy characteristics // Electronics and communication. — Thematic issue "Electronics and Nanotechnology". — [*Zhuykov V. Ya. Algoritm avtomaticheskoy klassifikatsii segmentov rechi na osnove avtokorrelyatsionnykh i energeticheskikh kharakteristik // Elektronika i svyaz'. – Tematicheskiy vypusk "Elektronika i nanotekhnologii"*]. – 2010. – No 5. – pp. 83–89. (in Russian)