Issues of speech technologies application in the Azerbaijani language

Rasim Alguliyev[1], Lyudmila Sukhostat[2]
[1,2]Institute of Information Technology, Baku, Azerbaijan
[1]*r.alguliev@gmail.com*, [2]*lsuhostat@hotmail.com*

***Abstract* — With the development of state-of-the-art technologies, there is a need for convenient and natural communication between a person and a computer. The literature on automatic speech recognition has used many methods for various languages, including machine learning methods for speech analysis and classification. In terms of speech technologies, namely, speech synthesis and recognition, speaker identification, dialect recognition, and speaker accent, little research has been done on the Azerbaijani language. This paper analyses the issues of using speech technologies in the Azerbaijani language. The emergence of computers with artificial intelligence that can understand the Azerbaijani language will contribute to a new experience and a comprehensive intellectualization of human life.**

***Keywords* — speech signal; Azerbaijani language; automatic speech recognition**

I. Introduction

Speech is a complex signal resulting from several transformations occurring at various information levels: semantic, linguistic, articulatory, and acoustic. Speech is the easiest and most convenient way to exchange information and the most natural way of interaction between people.

The automatic speech recognition system helps us with this. Such a system allows a computer to take an audio file or direct speech from a microphone as input and convert it to text, preferably in spoken language script [1].

The main speech technologies include speech synthesis, accent recognition, speech and language recognition, emotion recognition, gender detection, and speaker recognition (Fig. 1).

Recent advances in speech recognition create a dynamic environment as this technology allows to:

- access electronic services via voice without being distracted by the screen of the control device (keyboard, mouse, etc.);
- increase the security of personal data by using voice authentication;
- simplifies the input of search queries;
- make web resources accessible to people with disabilities.

Human speech, dialects, and accents have tremendous variety, and this variation in speech patterns is one of the biggest hurdles in building an autonomous speech recognition system.

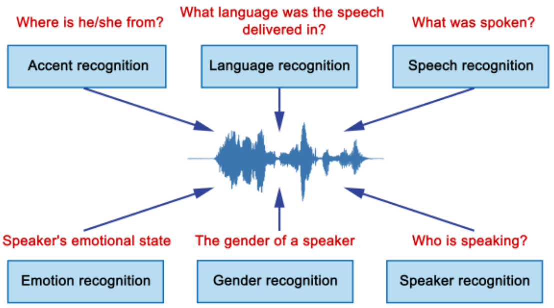

Fig. 1. The different speech tasks.

Furthermore, the problem arises when we add various factors like gender, style, and speed of speech. Another obstacle to creating an automatic speech recognition system is the presence of a large speech corpus, particularly for the Azerbaijani language. Along with the use of various methods and algorithms, it is necessary to consider the features of natural language in developing these systems. In this regard, at the moment, there are only some learning models for a small number of languages [2].

II. AUTOMATIC SPEECH RECOGNITION

As one of ICT, modern speech technologies have unique capabilities and are widely used in such areas as e-medicine, forensic phonetics, distance learning, and others.

Speech technologies using Big Data, the Internet of things, and cloud technologies contribute to convenient and natural communication between people and a computer. Various methods have been proposed to solve the problems of automatic language identification based on machine learning methods, including deep neural networks [3].

The main stages of state-of-the-art automatic speech recognition systems are shown in Fig. 2.

A. *Feature extraction*

The feature extraction process is used to remove unnecessary information from the signal. A good feature extraction algorithm should be able to extract features in real time and contain as much information as possible. Feature extraction algorithms can also be classified based on speech features into temporal and spectral features. Time analysis methods analyse the audio signal in its original form in the time domain. The spectral analysis uses the spectral representation of a speech signal, i.e., the frequency domain. Some of the techniques used for feature extraction are MFCC (Mel-frequency cepstral coefficients), PLP (perceptual linear prediction), DWT (discrete wavelet transform), Relative Perceptual Spectral Linear Prediction (RASTA-PLP), and LPC (linear prediction coding).

B. *Classification*

After extracting the features, they are passed to the classifier's input. The task of the classifier is to study the relationship between the given sound input characteristics and the corresponding text. First, they are trained using a large enough dataset to recognize specific patterns in the speech signal. The most commonly used classifiers are HMM (hidden Markov model), ANN (Artificial neural network), and SVM (support vector machine) [3].

C. *Language model*

An efficient language model is needed to create a real-time automatic speech recognition system. It uses the structural constraints of the language to predict the probability of a word occurring for a particular sequence of words. The difference between a classifier and a language model is that the classifier matches speech signals to the closest possible sequence of words. In contrast, the language model tests the probability of occurrence of the sequence of words generated by the classifier. Since in all world languages, some phrases sound the same but with different meanings. It can be fixed using a language model in conjunction with a classification model.

D. *Acoustic model*

Acoustic modeling analyses the training data in terms of relevant characteristics, for example, using various possibilities, expressing them as probabilities, and then combining these probabilities into a model, for example, HMM [4]. Various types of acoustic models have been used for automatic speech recognition, such as GMM (Gaussian mixture model), DNN (deep neural network), and LSTM (long short-term memory) in combination with HMM, and CNN (convolutional neural network) [5].

E. *Decoding*

Feature vectors are decoded into the linguistic units that make up speech using acoustic models derived from the recordings and their corresponding transcripts. Linguistic and pronunciation knowledge is often used to improve decoding performance.

III. Problems of Automatic Speech Recognition

Automatic speech recognition systems differ in the type of speaker, style of speech, and vocabulary size.

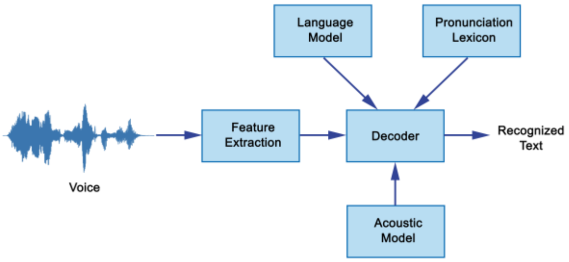

Fig. 2. Typical automatic speech recognition process.

They can be speaker-dependent, speaker-independent, and speaker-adaptive (Fig. 3). The efficiency of an automatic speech recognition system's efficiency depends on the dictionary's size.

The main problems of speech recognition are as follows:

- the same speech fragments have different characteristics and different duration;

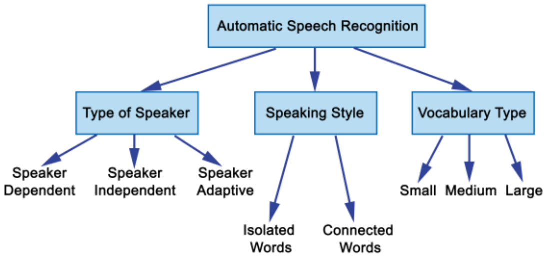

Fig. 3. Speech recognition systems classification.

- faulty and unclear pronunciation;
- poor diction of the speaker;
- high level of extraneous noise;
- insufficient or poor training of AI models;
- significant similarity of dictionary words;
- pronunciation with different intonation;
- the grammar used by the speaker and received by the system, the position of the microphone, and the user's speech speed;
- accent and dialect of the speaker.

IV. Speech Technologies in the Azerbaijani Language

One of the main challenges for state-of-the-art research on speech production and speech technologies is the understanding and modeling of individual variations in spoken language. People have their speaking styles, depending on many factors. These differences demonstrate the difficulty of modeling large-scale speaker-independent systems designed to process information in a particular language. People learn over the years, to some extent, to identify and interpret most of these speech aspects.

Over the past few decades, significant progress has been made in automatic speech recognition for various languages based on a given speech pattern [6, 7].

The Azerbaijani language belongs to the Turkic group of languages. This language group has more than 50 languages.

Azerbaijan is currently developing automatic speech recognition systems using the most modern methods [2, 8, 9].

Deep learning is fast becoming the method of choice over traditional automatic speech recognition methods. In addition, most research is developing toward multimodal and unsupervised speech recognition and Natural Language Processing (NLP) [9, 10].

Conclusion

Voice recognition of the language ensures its entry into all modern machine learning and artificial intelligence systems. Recognition, processing, and translation of speech require complex engineering and linguistic solutions. This paper considered the issues of using speech technologies in the Azerbaijani language.

Summing up, the following priority areas can be identified:

- Development of large speech corpora.
- Development of continuous speech recognition technologies, including ready-made acoustic-phonetic models.

- Creation of a program complex for restoring distorted and noisy speech messages in both a limited (thematic) dictionary and a mixed one.
- Creation of interpreters that correctly convey the meaning of a highly distorted speech message.
- Automated detection of linguistic information in texts and speech messages for biometric control.

References

[1] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: A survey," Multimed. Tools Appl., vol. 80, pp. 9411-9457, 2021.

[2] K. Aida-zade, S. Rustamov, E. Mustafayev, and N. Aliyeva, "Human-computer dialogue understanding hybrid system," in Proc. IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), July 2012, pp. 1–5.

[3] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, and M. Masud, "Spoken language identification using deep learning," Comput. Intell. Neurosci., vol. 2021, Article ID 5123671, pp. 1-12, 2021.

[4] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, Vol. 77, pp. 257-286, 1989.

[5] A. Pervaiz, F. Hussain, H. Israr, M. A. Tahir, F. R. Raja, N. K. Baloch, F. Ishmanov, and Y. B. Zikria, "Incorporating noise robustness in speech command recognition by noise augmentation of training data," Sensors, Vol. 20, pp. 1-19, 2020.

[6] M. Biswas, S. Rahaman, A. Ahmadian, K. Subari, and P. K. Singh, "Automatic spoken language identification using MFCC based time series features," Multimed. Tools Appl., pp. 1-31, 2022.

[7] Y. Zhang, et al. "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," IEEE J. Sel. Topics in Signal Proc., Vol. 16, pp. 1519-1532, 2022.

[8] A. Abbasov, R. Fatullayev, and A. Fatullayev, "Speech technology market in Azerbaijan," American J. Manag., Vol. 21, pp. 95–101, 2021.

[9] M. Mahmudov, R. Fatullayev, S. Abbasov, A. Fatullayev, and N. Abdullayev, "NLP systems for the Azerbaijani language and theoretical and practical tasks of creating the national corpuses," Turkology, No. 4, pp. 15-28, 2016.

[10] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," IEEE Access, Vol. 7, pp. 117327–117345, 2019.