Speech Signal Processing for Low Resource Languages

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

ASR in natural language and corresponding machine learning methods are essential for research nowadays. The purpose of this work is to review the main challenges and approaches in Automatic Speech Recognition (ASR) for low resource languages in the scope of speech segmentation, automatic feature extraction and feature generation.

Keywords*:* ASR, speech features, clustering, machine learning

Introduction

Over the past decades, there has been significant development of the machine learning paradigm. This paradigm is often used for automatic speech recognition, ranging from mobile devices to "smart home" and ending with space systems [1, 2]. The most effective and commonly used machine learning techniques are artificial neural networks, support vector machines [3], Gaussian mixture models [4], and hidden Markov models [5]. Recently, the deep learning approach [6] has also received a significant expansion, which has improved the efficiency of speech signal recognition.

Although commercial speech recognition systems based on the machine learning paradigm have been available for some well-defined applications such as dictation and transcription, there are still many unsolved problems in this area. These include recognition in a noisy environment, multimodal recognition, multilingual recognition, recognition in an unknown language [7].

Modern automatic speech recognition systems have achieved acceptable accuracy results and are operated widely in various fields. Further improvement in accuracy will further expand the scope of speech technologies in the daily life of ordinary users. However, standard supervised recognition methods [6] have essential obstacles for further improvement in recognition quality due to some inherent limitations.

The most proven standard methods are based on machine learning algorithms with well-prepared training data sets (labeled speech corpora) and using them (hidden Markov models, neural networks, etc.). In such systems, recognition accuracy is limited by the training data set's volume, balance, and labeling accuracy. Often such systems have difficulty recognizing new words and rarely use contextual combinations of known words. There are several commonly recognized problems in modern speech technologies, the solution of which will contribute to their transfer to a new qualitative level. One of these problems is the problem of detecting unknown words in a speech signal [8]. The ability of the system to independently detect an unknown expression in a speech signal and learn to recognize them is a valuable feature that improves the quality of recognition. Modern designs are still terrible at this, and the presence of this feature will change the approaches used for acoustic modeling. Instead of simply accepting only what we already know, we need the ability to determine the fact that we do not know something. Before recognition, it is necessary to automatically divide and classify the signal into known and unknown fragments (words, phrases). The basic speech pre-processing pipeline is shown in FIGURE 1. The essential research interest is concentrated on advanced feature extraction and segmentation approaches, as these parts may potentially decrease the workload on speech recognition models. The work intends is study the possibility of creation a speech recognition system that are adaptable and flexible by posing a very extreme situation when it is necessary to learn the entire language without having any knowledge about it [1]. Due to resource limitations, it is reasonable to consider as the highest level of the language structure for the unsupervised learning system only the lexical level. Speech and linguistic technologies accompanied with algorithms that can complement systems with preliminary training in a

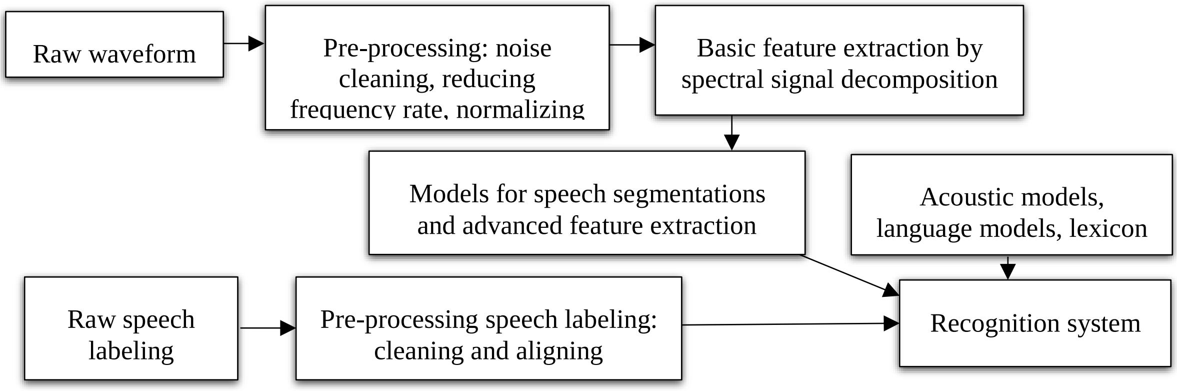
Raw waveform
Pre-processing: noise cleaning, reducing frequency rate, normalizing
Basic feature extraction by spectral signal decomposition
Models for speech segmentations and advanced feature extraction
Acoustic models, language models, lexicon
Raw speech labeling
Pre-processing speech labeling: cleaning and aligning
Recognition system

FIGURE 1 ASR pipeline

barely manually annotated corpus (low-resource languages or dialects) [8]. In addition, it allows identifying new and unknown words not included in the training data set in the speech stream. Fast unsupervised learning of recognition algorithms from a minimum data set is still a little-studied problem; some competitions are held annually [9].

Hence it remains challenging problem to extract effectively the information about the structural linguistic units of an unknown speech signal in an unknown language without prior training. Information about structural linguistic units is retrieved sequentially from low levels to higher ones. The lowest-level units are individual micro-local segments of the signal at the level of individual acoustic events, and the highest-level unit is subphonemic units of individual words. The process of analysis and transcription is performed by the sequential iterative clustering of low-level units to aggregate them into higher-level units up to individual words; density methods are used to cluster with a previously unknown number of clusters [10]. For these purposes, for each level, specialized clustering algorithms are developed that take into account the linear structure of the speech signal and the contextual features of the distribution of various types of segments at this level. One of the possible approaches for evaluating the solution's quality to this problem is modifying the method of dynamic adjustment of the time scale called Dynamic Time Warping (DTW) [11]. The main focus of this work is on robust algorithms for multilevel segmentation of the speech signal [12] and automatic classification of the received segments. The primary accented has been given to methods that allow extracting essential information about the structure of a raw speech signal in an unknown language using unsupervised learning. It allows to use unsupervised learning algorithms for an unknown language by identifying stable linguistic units in this language at various levels and their subsequent transcription. Transcription of speech signals means its encoding in the form of a compact textual description about the structure and various properties of the analyzed signal, taking into account its pronunciation. Transcription can be carried out hierarchically at all considered linguistic levels [2]. The machine learning system is immersed in an unknown language environment; hence, due to the generalization of the processed speech information and the analysis of the patterns, the system independently learns to identify various language structural units in the speech stream, up to individual words. This problem remains difficult to solve at the machine level, where the dominant paradigm is massive learning using large datasets manually labeled by humans.

Hence, it is essential to focus on 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. One of the main advantages of approaches in [9, 11] 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 best of their features. We hope that the provided survey will encourage the readers to participate in facing the challenges in the broad field of speech recognition.

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, 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

[2] Lu L, Ghoshal A, Renals S Cross-Lingual Subspace Gaussian Mixture Models for Low-Resource Speech Recognition 2014 *IEEE-ACM Transactions on Audio Speech and Language Processing* **22**(1) 17-27

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

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

[5] 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

[6] Cortes C, Vapnik V Support-vector networks 1995 *Machine learning* **20**(3) 273-297

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

[8] 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

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

[10] Nguyen G et al Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey 2019 *Artificial Intelligence Review* **52**(1) 77-124

[11] Gonzalez-Dominguez J, et al. A real-time End-to-End multilingual speech recognition architecture 2015 *IEEE Journal of Selected Topics in Signal Processing* **9**(4) 749-759

[12] Mitra V, Nam H, Espy-Wilson C, Saltzman E, Goldstein L Articulatory Information for Noise Robust Speech Recognition 2011 *IEEE Transactions on Audio Speech and Language Processing* **19**(7) 1913-1924