# 4.5.29. THE OVERVIEW OF ON-LINE SPEECH PROCESSING ALGORITHMS FOR SPEAKER IDENTIFICATION SYSTEMS

Korshakov A.V.
Russian Science Center “Kurchatovsky Institute”, Moscow, Russia

The presented paper is an overview of variety of speech recognition methods existing today for using in speaker identification systems, which in turn may be part of a large variety of access control systems or security systems. It is a large amount of publications on the matter released lately and the number still increasing. This, we believe, deals with the potential preeminent practical interest in usage of devices performing a fluent speech recognition function, voice recognition function and/or acting accordingly with the vocal commands. The possible practical implementation varies from simple “vocal typing“ or implementation in robotics to complicated “antiterrorist” voice identification systems as a common and/or special security measures. The firsts are already not far from reality, while the seconds in most parts are still remains a subject of “near science fiction”. Nevertheless, there are certain facts, making us to believe in possibility of constructing reliable (!) voice recognition and identification systems in the nearest future. The prerequisites of this possibility and such facts itself are follows in the paper. From this point and so on (unless told to be opposite) the word “algorithm” used in a wide meaning which includes, for example, method, a set of algorithms, or even a software product as a hole implemented particular processing algorithms.

It is needles to say that a big progress in this field of science have been made since the beginning of research on artificial intelligent, but this process is still far from completion. At the beginning a huge steps were made due to foundation of classical linguistic. But then success was tempted by the insufficien of computational power of computers available at the time as well as by not full understanding of such process as “speech” it self. At the same time and little later vast side knowledge was injected in the field, including processing with the artificial neural networks, and various algorithms based on hypothesizes of how human brain functional during speech processing such as the results of fMRI investigation of person performing a speech tasks.

Today, there is several levels of “recognition” of speech can be noted. Acoustic level or level of phonemes deals with separation incoming speech to independent language-specific phonetic units. Morphological level or level of words performing interpretation of words using information on incoming phonemes acquired at the acoustic level. Next is semantic level. The algorithms of this level trying to interpret the meaning of words. Next one is (or 2 composed levels) is sintagmatic and syntactic levels deals with complete logical language unit and/or with sentences. The last one closely interact or sometimes ultimately merge with the previous is probably most complicated and seems to be not yet can be fully performed by the machine at the time deals with the prosodic variance of a speech process such as intonational or even sarcastic notes in speech. Due to the fuzzy nature of human natural language the borders between these levels are blur and one level is difficult to completely separate from the other. For example sometimes phonemes level information proves to be ambiguient and information from one or even several upper levels is needed to recognize and interpret incoming sound correctly. Each one have its’ own problems and currently no optimal solution had been found.

The paper provides reader with a brief classification of existing algorithms of speech processing providing a man-machine interfaces mainly. Such as separate them to 2 big classes – speaker dependent and speaker independent algorithms. First ones must be trained (in usage of neural networking term, but artificial neuron networking not presumed every time) for one's speaker voice and then can be use by the same speaker with a relatively high results compared to the second class. Such algorithms also provide a voice recognition or identification of the owner by the voice function. The second ones have an advantage of "correct" recognition of speech indiscriminately to speaker but also they are usually far more sophisticated and complicated in usage and in preliminary tuning and as result some error levels is common. Nevertheless definitely those last will be methods of choice when it comes to universal usage.

The paper mainly concerns the first level of recognition from the list above and its interaction with the last one. Most common problems which arise during phonetic speech processing for recognition are described as well as most vital conditions of sound signal perception. This primarily concerns an attention focus of speech perception process and influence on noises. Eventually perception it self can be acknowledge as an independent problem ("cocktail party problem" and set of independent components analysis methods used to solve it is described). Speech signal is a quasi periodic process with spikes and silence areas or "events" are includes in it which marks an essential parts in speech. Those marks usually can be assigned to particular phoneme. Therefore it is impotent to be able to identify such marks. This problem usually solve by using different ways of spectra and cestrum of a signal analysis, with using curtain signs to determine varies in different works and in a different approaches. Using a spectrographic representation of signal are also usual. Most phonemes have there own spectral signature, yet not unique enough to distinguish them definitely. Still phonemes can be distinguished at some degree and classified at some classes. Classification and clusterization problem are also occurred during the phonetic analysis of speech. It is usually arising when it is necessary to separate sound thread to a base phonetic sequences (or phonemes) and thou associate them with words sequences and meanings respectively. It is rather difficult to accomplish such task at the 100% level. Many of sound that we used in an every day speech have a trend to combine with each other (composing the "mean among two" sound), or have a fuzzy borders, which makes the separation task very complicated. There are many methods to perform such classification. Among them are those which using features of particular nature human language, or those which based on abstract models like "Hidden Markov Chains" etc, or based on complicated lows of correlation dependences among different marks in a processed speech signal. Artificial Neural Networks are also play a great role to solve this. Such algorithms are often called "intellectual". And for the reason that speech is an intellectual process it self, it is obviously that the particular algorithm also need to be intellectual by nature. And It must based on some fundamental principles of speech generation (at a phonetical level and at other levels as well) by the human being. Seems to be trying to understand algorithm of speech signal analysis in time-frequency domain in a human brain and "in a way" try to mimic it is a most fruitful approach.