Some problems in automatic identification of spoken utterances and speakers

Ramishvili, G.

journal: STL-QPSR
volume: 11
number: 2-3
year: 1970
pages: 036-040

http://www.speech.kth.se/qpsr
III. AUTOMATIC SPEECH RECOGNITION

A. SOME PROBLEMS IN AUTOMATIC IDENTIFICATION OF SPOKEN UTTERANCES AND SPEAKERS

G. Ramishvili*

This paper describes two main problems which are under development at the laboratory of transformation of information at the Institute of Control Systems in Tbilisi. These are automatic recognition of speech commands and automatic identification of speakers.

Automatic recognition of words

In 1963, A. G. Kakawridze, who is the head of the department, constructed a special device that recognizes 20 oral numerals and some other commands directed to a printing machine and to a mobile automaton. The spectrum analyzer used in this device included seven band-pass filters. The output of each filter after rectifying and smoothing was quantized on three levels. The channels were synchronously sampled after every 20 msec. Some steady and relatively invariant spectral positions were selected for every command and the definite sequence of these was included in the logical set of decoding elements. The accuracy of this device was independent of loudness and duration of the utterances, but was very sensitive to different speakers.

Some possibilities of word recognition by means of a computer were studied later on. More complicated programs of acoustical description and learning process were then adopted. In the simplest case of the decision function program it was admitted that any class of recognizing words can be represented by a hypersphere with equal radii, so that they can be described by one point in the center. In further developing the programs it was taken into account that the points in the space of the description are not uniformly distributed.

In the last case the areas of the classes were described by means of calculation of some points of gravity, accomplished in a learning procedure.

Let us suppose, for instance, that we have $n$ learning utterances of a class and we want to select such $m$ ($m \ll n$) which sufficiently well characterizes the boundary of this class. One of the ways is to prove all $\binom{n}{m}$

* Guest researcher at the Dept. of Speech Communication April 14-May 15, 1970.
combinations and select such m points that provide maximal sum in one of the edges of the multidimensional graph. When the n is large, it is better to count distances between utterances in pairs, selecting the largest ones in any descriptive dimensions of a class.

Such selection of the "frontier parts" of classes is favorable if the areas of the classes do not overlap. If they actually do the local subclasses with their own center of gravity have to be introduced. For this purpose an algorithm was foreseen which builds additional subgroups for the (n-m) utterances which remained after building the main group of m of them. Such complex algorithm gave the possibility to recognize about 50 words with a high accuracy.

On the other hand, we have seen that the automatic recognition of words in analogous devices as well as in computer programs works much better for one person than for many speakers. Preliminary identification of the voice, for normalization purposes, should therefore reduce several difficulties in working with many speakers to the ideal situation of a small group with similar voices.

Automatic recognition of speakers

The problem of automatic identification of speakers by means of individual characteristics of the voice arose in our laboratory in close connection with the problem of speech recognition.

It is well known that the information in the speech wave about individual voice characteristics appears simultaneously with the semantic information so it is very difficult to separate them for independent use.

The general progress in solving this problem depends very much on our possibilities to operate quantitatively with the meaning, as well as with many kinds of prosodic information about intonation, loudness, rate of speech, and so on. All of them being mixed in a signal create supplementary difficulties in objectively classifying voices. One of the simplest means to avoid these difficulties is to consider determined semantic segments of speech (determined syllables or words as representatives of an individual speaker's voice). However, this assumption overlooks the great influence of context on the spectral patterns of words.
The second approach to voice identification is based on some long-time statistical parameters of the speech signal considering the whole of it as a stationary random process. The reason for this approach is the fact that the individual properties of the speaker's voice affect the listener's hearing constantly during the whole speech and can therefore be considered as some constant background if the utterance is large enough. Now, if we take the existence of some laws of the language into consideration, especially the appearance of the probability of occurrence of phonemes in speech, we can expect the stabilization of the message statistics after some duration of the piece of speech for all speakers, all the differences concentrating in their individual manner of speaking.

In two experiments described below the here mentioned integral properties of the speech signal have been used. As descriptive features of spoken utterances we took the intervals between the zeros and the extremal points of the speech wave. Such parameters are simple to obtain and they are not very much dependent of the loudness of the pronunciation. Thus, it is not necessary to normalize the signal by its amplitude.

By means of a special discriminator consisting of 18 channels we can represent the speech signal as a distribution of probabilities of intervals of the different durations in the speech wave.

It was found that the form of this distribution does not depend on the semantic contents of speech and it reflects mainly the characteristics of the voice inside the same language. It is fair, however, if the duration of speech is not less than 10-15 sec. Below 10 sec the piece of speech cannot represent significantly the statistical properties of the language.

The output of the equipment may be described

\[ X(\tau_k) = \frac{\tau_k n_k}{\Delta \tau_k T} \quad k = 1, 2, \ldots, 18 \]

Here the \( \tau_k \) is the middle interval of the k-channel, \( n_k \) is the counted number of intervals gated in the k-channel, \( \Delta \tau_k \) is the width of each channel, and \( T \) is the duration of the analyzing speech signal.

\[ T \approx \tau_k n \]

In the testing experiment each person pronounced an arbitrary test. The identification was carried out on a computer by calculating Euclidian distances between the realizations of the descriptions and the modes.
discrimination of 20 male voices by means of zeros and especially extremal points gave us 90% correct identification.

Next experiment was based on the spectral analysis of the speech wave. The speech signal $S(t)$ was analyzed by means of 24 quarter octave band-pass filters. The output of each filter was rectified, smoothed, and integrated on the whole time $T$ of the spoken utterance.

The output of each filter was described as

$$c_i = \frac{1}{T} \int_{0}^{T} \int_{0}^{t} S(\tau)g_i(t-\tau)d\tau \ dt$$

Such an approach gave us very similar results of identification.

The same integral method of representation of some particular speech parameters was investigated. For instance, it was experimentally shown that the pitch is a good parameter for characterizing the individual voice when it has been following along a large time interval in a speech utterance. Estimating some dynamic parameters of the signal, for instance, the average value of the rate of the pitch, the rate of the moving formant frequencies, and so on, gave us promising results too.

A useful information has been obtained also in experiments with frequency-limited speech and with its limitations by the amplitude. Some rules were studied about changing the nature of the signal by compressing it from high and low frequencies, as well as by clipping the real signal and the high-frequency modulated one, in order to transmit the information about the speaker.

The future development of the investigations has to be augmented with studying individual properties in every step of speech production, as well as dynamic properties of speech in their phonetic, phonological, and linguistic aspects. I hope that the experimental material obtained in the Dept. of Speech Communication, KTH, will serve to attend some of these goals.

Acknowledgments

I would like to thank Prof. Fant and all the members of the staff at the Dept. of Speech Communication for their hospitality, their attention and care to me during my time in Stockholm.

refs., see next page
References:


