Vowels Detection / Recognition on the Base of Short Cross-correlation Function Side Peak Parameters.

W. E. Antciplov

Institute of Radioengineering and Electronics of Russian Academy of Sciences, Moscow, Russia

antciplov@cplire.ru

Abstract

The report presents the latest results in continuous speech phonetic analysis, concerning the problems of stable and effective speech recognition. These results are obtained on the base of earlier discussed [4] approach. At the heart of approach underlies the analysis of dynamics of short cross-correlation function (CCF) parameters, namely the time location and the value of CCF side peak. This approach in natural manner gives rise to the so-called detection / recognition paradigm in phonetic speech processing. From the general point of view the paradigm implies the detection of statistically homogeneous signal fragments and their structure clarification [5]. Both detection and recognition procedures are in frames of approach mutually dependent and represent two sides of the one uniform processing. The effectiveness of technique proposed is illustrated by a number of real speech processing examples.

1. Concept

Cross-correlation analysis as a special case of time domain speech processing methods though is not as popular as numerous spectral ones, but as it is shown below, allows solving some problems very efficiently. Such paradoxical situation is conditioned by the nature of speech signal: being non-stationary as a whole, it has noticeable quasi-stationary fragments. These fragments can be formed by vocalized sounds on the one hand, and by the noise-like sounds on the other. Time intervals of quasi-stationarity have as a rule the duration from units up to dozens of base tone periods, i.e. constitute 30 - 100 ms. It implies, that to utilize correlation analysis in speech processing, we should limit ourselves to only short cross-correlation functions (SCCF).

Normalized SCCF of noise-like sounds has a form of a delta function, so its characteristic property is the only one, narrow, unit maximum at the time origin. Vocalized sound represents almost periodic signal whose normalized SCCF is also periodic. This implies that its SCCF has side maxima which retrace the central one, i.e. there are almost unit side peaks spaced by the base period (5-10 ms) from the origin.

This simple consideration leads to main idea of approach – to monitor the values and time locations of side peaks of current SCCF along the speech signal. Having peak dynamics we can analyze it for the purpose of detecting extrema in its behaviour and for the matching this characteristic time moments with some classes of phonems – maxima with the vocalized, minima with noise-like. Fig.1 presents the example of such monitoring.

It is worth to note that all intermediate SCCF and side peak parameters calculations that are used on phase of detection are utilized further in recognition phase. So it gives that the approach became uniform and self-consistent and gives rise to the detection / recognition paradigm in phonetic speech processing.

2. Technical details

The technique of vowels detection, based on the dynamics of CCF side peak value, was described earlier [4]. So let us make the only additional remark about detection procedure. In view, that detected characteristic time moments correspond to the signals most correlated behavior, time fragments surrounding such detected moments are as far as possible stationary and represent the time intervals where local spectrum analysis is in the best way valid, i.e. the best place for spectrum recognition. By the way, near the characteristic time moments the almost stationary behavior of other parameters, for example formant frequencies, is also observed and as discovered by other researchers [6], such intervals (of formant frequency trajectory inflection) are most informative for vowels recognition.

Taking into account the above remark, the vowels recognition technique was based on almost, up to the realization nuances, classical spectrum analysis. First of all, in view, that SCCF calculations were previously made, it is convenient to estimate spectrum by spectral density function which is the counterpart of a SCCF in frequency domain (or simply the correlogram). By the way, the correlogram already includes the required averaging. Second, because all the informative frequencies are greater than the base tone, it is enough to accept for Fourier transform of CCF the symmetric time interval laid from the left and up to the right side peak. Technically it was
made by windowing CCF by, for example, Hanning function whose duration is supposed to be equal to previously estimated time location of SCCF side peak. Third, because SCCF is, at least theoretically, implied as symmetric about the origin, Fourier transform degenerates into cosine transform that was practically used.

The last nuance concerns the choice of distance between frequency samples. Such a choice must be made on the base of the following compromise: on the one hand the distance must be small enough to prevent the loss of significant information, on the other hand it is desirable that the redunancy among the sampels would be as small as it is possible, i.e. the sampels must be distributed far enough one from another. We accept the distance equal to the estimated base tone which is the value inversely proportional to the time location of SCCF side peak. The substantiation of such choice become clear in view that source speech signal (vocal cords oscillations) is exactly a set of harmonics whose frequencies are divisible to the base tone (fundamental frequency).

Estimated in such a way the spectrum contains ~ 3kHz / 150Hz ~ 20 sampels. Let us notice, that this information quantity is approximately equal to that, which were supplied by other feature extraction methods, for example LPC, per each analyzed frame. The advantage of our approach lies in a couple of aspects. First, there is no need for features for every adjacent frame, the set of which covers with overlapping the whole of the signal duration time interval. As mentioned above, we form the spectrum only for detected characteristic time moments the number of which is substantially less than the number of frames. Second, estimated spectrum samples (in comparison with LPC features for example) do permit explicit interpretation in terms of speech production peculiarities: formants frequency shifts, Q-factors, energy etc. Because of a huge phonetic data on this subject, the vowels recognition became a well-defined procedure.

3. Experimental results

To illustrate the effectiveness of the technique proposed we present below a number of real speech processing examples. All the speech samples relate to isolated vowels male pronunciation and have the following properties: sample rate 22 KHz, sample size 8-bit. The speech samples coupled with their classification were adopted from Section of linguistique, Universite de Lausanne site [7] devoted to production of the sounds of the language and the principal divisions in which the system of the International Phonetic Alphabet (I.P.A.) classifies them.

The descriptions of the vowels below are grouped according to the following principle: an initial classification is made based on degree of aperture that corresponds to the distance between the palate and the tongue's highest point, within each such group, the vowels are then divided according to mouth shape that is determined by the general position of the tongue in the mouth - front vowels (tongue body in the pre-palatal region), back vowels (tongue body in the post-palatal or velar region) and then as rounded (labial resonator active) or unrounded (no labial resonator/no labial resonance).

3.1. Close Vowels (First Degree of Aperture)

3.1.1. Unrounded close front vowel

Fig.2. Spectrum of phoneme “I”

3.1.2. Rounded close back vowel

Fig.3. Spectrum of phoneme “U”

3.2. Half-Open Vowels (Second Degree of Aperture)

3.2.1. Unrounded half-open front vowel

Fig.4. Spectrum of phoneme “E”
3.2.2. Rounded half-close back vowel

Fig. 5. Spectrum of phoneme “O”

3.3. Open Vowels (Fourth Degree of Aperture)

3.3.1. Unrounded open front vowel

Fig. 6. Spectrum of phoneme “A”

All graphs shown represent the normalized logarithmic spectra in the band of 0 – 4 KHz of frequency domain.

4. Discussion

As it follows from the above examples the discussed approach gives reasonable results in vowels recognition, which are in a good agreement with the known phonetic data [8]. First, one can see that with increase in degree of phoneme aperture the band of significant first harmonics becomes wider and its centre of gravity (first formant) shifts to higher frequencies. It corresponds to the known fact that open vowels have the higher consequent formants than closed ones. Second, the data given reveals the difference in rounded / unrounded opposition. Indeed, one can see that unrounded spectra have more noticeable maxima in 1-2 KHz band (second formant) than the rounded spectra do. This fact can be explained if we consider the labial resonator as some low-pass filter whose cutoff frequency in order of value is equal to its first resonance. Without labial resonator all mouth shape resonances (formants) take place, when labial resonator is active its first resonance moves to ~ 1 KHz (it is possible in very irregular resonators) and the formant near its cutoff is suppressed. Third, the difference in front / back vowels production opposition appears seemingly in slight frequency shift of formants from their position for central vowels – for front in low, for back in high directions.

5. Conclusions

The discussed above detection / recognition paradigm presents in some extent new, not traditional approach to speech recognition. As it is shown above, this approach reveals on the one hand known in speech processing features on the other hand some new aspects. In addition it demonstrates fair agreement with the known phonetic data concerning speech production. In spite of still existing scepsis with respect to adequate, computer-based acousto-phonetic speech recognition, the results obtained give hope that the opportunities in this field are yet far from their limits.

6. Acknowledgements

I would like to thank my colleague Vladimir Morozov for the fruitful discussions about approach engineering details and results obtained. I am also grateful to Nicitov Sergey acting director of IRE RAS for long-term support of investigations in the field of speech recognition.

7. References