Speech Analysis and Recognition Synchronised by One-Quasiperiodical Segmentation

Taras K. VINTSIUK, Mykola M. Sazhok
NAS Institute of Cybernetics & UNESCO/IIP International Research-Training
Centre for Information Technologies and Systems
Kyjiv 252022 Ukraine

Abstract. It is shown that the best ASR results are attained when a pre-processing is carried out synchronically with pitch. Specifically, an analysis step has to be equal to the current one-quasiperiod duration and current analysis intervals have to consist of an entire number of quasiperiods with total 45-60 ms duration. Quasi-periodicity and non-qusiperiodicity models and measures as well as their applications for the optimal segmentation of speech signals into one-quasiperiods are given and discussed. Then the ways to embed these pre-processing results into the recognition procedure are described.

1 Introduction

A lot of problems in speech signal pre-processing still await for solution. Among them there are such questions: Is it really necessary to perform speech signal pre-processing before its recognising? If so, than must it be synchronised by pitch or not? What is the analysis interval duration? And what is the analyser on the whole?

In this paper it is shown experimentally that speech signal pre-processing, if it is performed before the recognition, must be fulfilled synchronically with a current pitch period. So analysis interval bounds must match the bounds of quasiperiods, and current analysis interval duration must be in range of 10–60 ms and more.

Further, there are considered: models of the speech signal quasi-periodicity and non-periodicity, similarity measures, the algorithm for the speech signal optimal partition into quasiperiods, pointing their beginnings, the ASR procedure, that is synchronised with the one-quasiperiod speech signal segmentation.

2 Influence of Discretisation Effects, Analysis Interval Length and Analysis Step on the Recognition Accuracy

The most used analyser of a speech signal is following. At first, speech signal is divided into segments or analysis intervals with the constant step ΔT and duration $\Delta T'$. Then, such picked up speech signal segments are analysed. It is obviously, that under such approach speech analysis intervals are placed randomly relatively

to the beginning of a speech signal. In [1] it is shown, that the signal amplitude spectre distinctively changes if the analysis interval is shifted even by one discrete. Thus, the results of analysis depend on discretisation effects that means the randomness of the analysis interval shift relatively to the speech signal beginning.

At the time of speech signal analysis typically the overlapped analysis intervals are used ($\Delta T < \Delta T'$), and their duration is in range of 10-30 ms. Although, it is no clarity neither theoretical nor experimental here.

To make clear how discretisation effects, analysis interval duration and step influent on the recognition accuracy the two series of experiments were made.

The recognition training and proper recognition were performed accordingly with algorithms based on the Dynamic Time Warping [1]. The training sample consisted of 500 speech signals for 100 isolated words (5 realisations per word). The test sample was similar, namely 500 realisations, exactly 5 speech signals per word.

In the first set of experiments the dependence of recognition accuracy on discretisation effects was studied. It was considered 30 different analysis interval durations $\Delta T'$ under the fixed analysis interval step $\Delta T = 15$ ms. Since the analysis step was invariable $\Delta T = 15$ ms and speech signal discretisation step was equal to $\Delta t = 50$ μs, therefore due to discretisation effects, that is by the shifting of analysis interval by different number n discretes, $n = 1{:}(N\text{-}1)$, $N = \Delta T / \Delta t = 300$, it was received 299 additional test samples for each original one. Thus, actually test sample consist of 300·5·100 = 150 000 word realisations.

In the second set of experiments 16 different join non-overlapping analysis intervals ($\Delta T = \Delta T'$) were studied. Here the number N of additional realisations was changing.

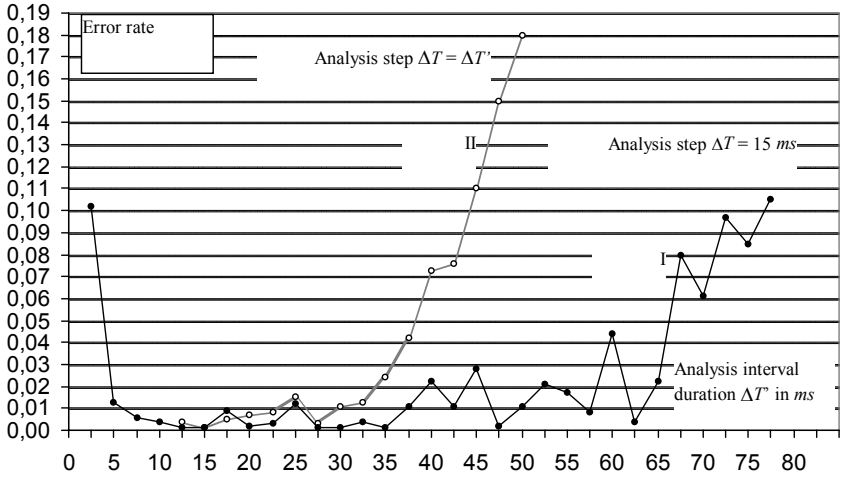

Fig. 1. The recognition error rate dependence on the analysis interval duration $\Delta T'$.

Speech signal analysis and similarity measures (both elementary and integral) used in the Dynamic Time Warping matching were based on the auto-correlation, co-variation, linear predictive, spectral, cepstral or coded descriptions [1].

In Fig. 1 the recognition error rate dependence on the analysis interval duration $\Delta T'$ under the invariable analysis step $\Delta T = 15$ ms (continuous curve I) and on the same analysis interval duration $\Delta T'$ under the condition $\Delta T = \Delta T'$ (dotted curve II) are given [2]. As a speech signal description it was used the 48-bit binary code, that was the discrete analogue of the auto-regressive spectre derivative sign on the set of 49 frequencies. As an elementary similarity measure for observed and reference elements it was used the Hamming distance [1].

Studying dependencies in Fig. 1 allows to conclude:

1) Curve I has emphatic oscillative quasi-periodical tendency with average speaker pitch period.
2) The smallest error rate ("cavities" on the Curve I) comes across when analysis interval duration $\Delta T'$ is fit by entire number of quasi-periods.
3) The best recognition accuracy is reached on the wide range of analysis interval duration from 10 to 65 ms. Sensitivity to one-quasiperiod synchronisation grows with the analysis interval increasing.

3 Speech Recognition Synchronised by One-Quasiperiodical Segmentation

3.1 Two-Level ASR System Structure

In Fig. 2 it is considered the two-level speech recognition system.

At the first level the problem of optimal current pitch period discrimination and speech signal partition into quasi-periods is solved. It consists in finding the best quasiperiod beginnings or the best one-quasiperiod segments.

At the second level the input speech signal marked out by one- I od beginnings is recognised.

3.2 Optimal Signal Partition into One-Quasiperiod Segments

In [3] the algorithm for optimal speech signal partition into one-quasiperiods is described. Each hypothetical one-quasiperiodical signal segment is considered as a random distortion of previous or following one taken with an unknown multiplying factor. The problem consists in finding the best quasiperiod beginnings under restrictions on both value and changing of the current quasiperiod duration and multiplying factor.

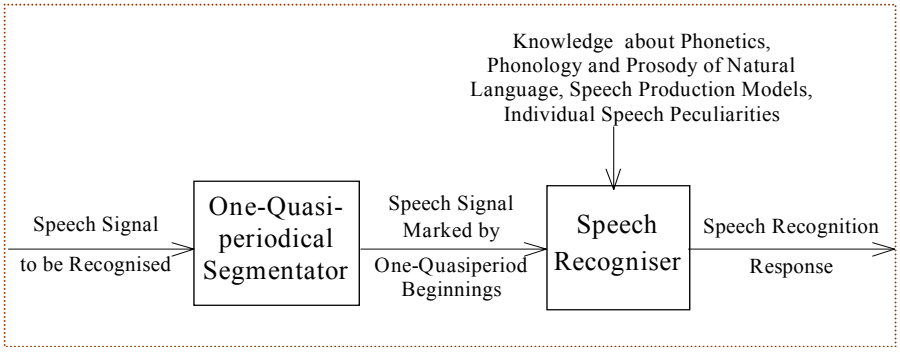

Fig. 2. Speech recognition system structure

Different elementary similarity measures for one-quasiperiodical speech segments comparison are introduced. Dynamic programming matching procedure guarantees a choice of the best speech signal partition into quasiperiods for which the integral sum of respective elementary similarity measures for hypothesised joint quasiperiods is the largest.

Thus, the notion of one quasiperiodical segment is applied not only to properly quasi-periodical signal segments, but it is extended to any kind of one too, particularly to noise segments.

In Fig. 3 and Fig. 4 the examples of the optimal speech signal partition into quasi-periods are given for quasi-periodical and non-periodical signals, respectively.

3.3 How to Run with Quasiperiods at the Stage of Recognition

When the speech recognition process we run with one-quasiperiodical segments by such a way. All one-quasiperiod beginnings are considered as the potentially optimal bounds of phoneme-threephones and each observed one-quasiperiodical segment is tested as a random distortion of reference one taken from the codebook of a so-called Speaker Voice File or Passport (SVP).

The individual SVP is computed through the speaker training speech data. Such parameters are to be estimated: a set (alphabet) of typical one-quasiperiodical segments, that is a codebook; acoustical transcriptions of phoneme-threephones in names of typical one-quasiperiodical elements; intonation contours for syntagmas. Thus, SVP describes the individual phonetic-acoustical diversity and peculiarities of pronouncing.

Further we will consider the ASR for words and phrases taken from the Word/Phrase-Book. Thereby, it will be not running with prosodic information.

Of course, SVP refers to general linguistic and phonetic data and knowledge base for a concrete natural language. Typical one-quasiperiodical elements are chosen from real speech training sample. They are specified by speech signal segments in time domain or by any other equivalent description, e.g. by linear predictive co-variance vector $e(k^1) \in E^1$, $k^1 \in K^1$ where K^1 is the name alphabet of

typical one-quasiperiodical segments and E^1 is a set of reference elements. We interpret k^1 and $e(k^1)$ as micro-phonemes or the first level speech patterns.

An observed one-quasiperiodical segment x is compared with a reference element e by such measure of similarity like

$$g(x, e) = \ln(1+(a-e(k^1))^T B\,(a-e(k^1))), \quad (1)$$

where x and B are the one-quasiperiodical segment given by its co-variance both predictive vector a and matrix B respectively.

Phoneme-threephone (PT) forms the second level speech patterns. The PT is the basic phoneme that is considered under influence of neighbouring phonemes in context, they are the first which precedes and the second which follows. For each natural language there are fixed about 2,000—3,000 PT. Each PT from SVP is specified by its speaker transcriptions in the individual microelement alphabet.

Since the isolated words or phrases recognition, the orthographic text of each hypothesised word or phrase is converted into phoneme and respective phoneme-threephone transcriptions. Accordingly to the latter and referring to acoustical phoneme-threephone transcriptions, a so-called initial model signal of the word or phrase, presented by a sequence of reference microelements, is composed. Then, in time domain, the non-linear transformations of the initial model signal are performed, and results of these transformations are compared with the input signal by using the Dynamic Time Warping procedure [1].

Non-linear transformations allow to repeat or remove a microelement (typical one-quasiperiod) between neighbouring ones. More exactly, it is forbidden to repeat the same microelement more than two times, and to remove than two microelements running. At last, the transformed one-quasiperiodic sequence has to have the same quantity of one-quasiperiodical elements as it is in the speech signal to be recognised.

Finally, running the one-quasiperiod-to-one-quasiperiod comparison accordingly with (1), the best integral matching for a hypothesised word or phrase is found.

Recognition response is the word or phrase with the best matching on a word/phrase-book.

3.4 Experimental Results

Two series of experiments were set. In the first one it was repeated the experiments described in the chapter 2. The difference was in that that analysis interval length and step were synchronised by one-quasiperiodical pre-segmentation. Any error

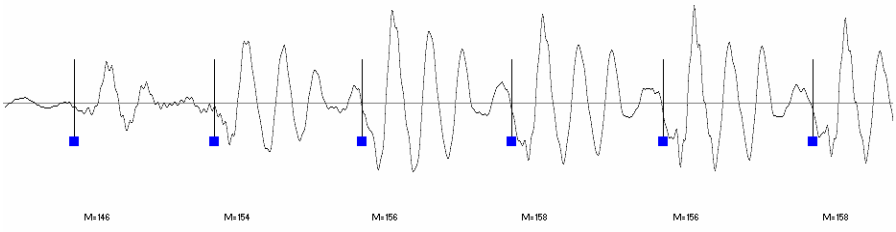

Fig. 3. One-quasiperiodical partition onto the voiced speech signal. The value of M shows the one-quasiperiod segment duration in discretes. Discretisation step Δt is equal to 50 μs.

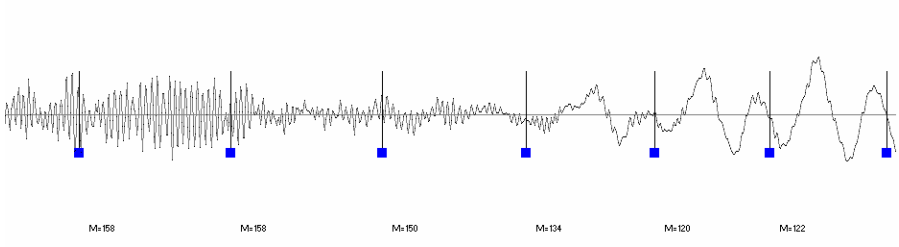

Fig. 4. One-quasiperiodical segmentation on the transitive noisy/voiced speech signal [su'].

4 Conclusion

In this paper by the experimental way it is shown that robust word and phrase recognition with higher accuracy is reached when the recognition procedure is synchronised by one-quasiperiodical pre-segmentation.

It is expected a similar effect in the automatic speech understanding under taking into account a prosodic information.

References

1. T.K. Vintsiuk. Analysis, Recognition and Understanding of Speech Signals. – Kyjiv: Naukova Dumka, 1987, 264 p, in Russian.
2. T.K. Vintsiuk, L.I. Khomenok. Influence of Discretisation Effects, Analysis Interval Length and Step to Speech Signals Recognition Accuracy. – Proceedings, 15 All-Union Seminar "Automatic Recognition of Sound Patterns", Tallinn, 1989, pp 81-82, in Russian.
3. Taras K. Vintsiuk. Optimal Joint Procedure for Current Pitch Period Discrimination and Speech Signal Partition into Quasi-Periodic and Non-Periodic Segments. – In "Text, Speech, Dialogue", Proc. of the First Workshop on Text, Speech, Dialog — TSD'98, Brno, 1998, pp 135-140.