Phoneme Recognizer Based Verification of the Voice Segment Type Determination

OLDŘICH HORÁK
Faculty of Economics and Administration, Institute of System Engineering and Informatics
University of Pardubice
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC
oldrich.horak@upce.cz

Abstract: - This document describes the verification of the result of voice segment type determination. The verification is performed by the Phoneme Recognizer. The voice segment type determination is based on the presence of fundamental frequency in the voice segment signal, and the several methods are used for the fundamental frequency presence demonstration. The fundamental frequency of speaker’s voice, which can be extracted from the voiced segment of the speech signal, is one of the basic characteristic features of the speech used in the speaker recognition process.

Key-Words: - fundamental frequency, signal processing, speaker recognition, voice signal, phoneme recognizer

1 Introduction
The information system’s user identification is the very important task of the information systems common security. One of the unusual biometric methods is speaker recognition. This method uses an extraction of the speaker’s voice tract anatomy parameters from the voice signal. The sound timbre of the speaker’s voice is dependent on and given by these parameters.

In the voice signal processing, the signal is divided in blocks – short segments of the speech with time duration in tens of milliseconds. The voice characteristic is based on features extracted from the segments. Some features can be extracted from the voiced segments only, other from surd segments. Therefore, the voice segment type determination is important for these tasks. [1, 2]

The fundamental frequency is one of the basic features of the voice. It is the common tone level of the speaker’s voice. This characteristic feature can be extracted from the voiced segment only. The presence of the fundamental frequency in the voice segment can be also used to determine the type of given segment. The detection of the fundamental frequency is sufficient to determination of the voiced segment, and the exact value of the frequency is not important in this case. [11]

2 Principles
The voice segment type depends on the voice signal part, which is cut from. Simplified, the segment from vowel is voiced, and the part of consonant signal is surd. It is only the approximation; the exact determination of the segment type has to be evaluated by a signal processing method.

2.1 Segment Type Determination methods
There are more methods to distinguish the voice segment type:

2.1.1 Fundamental Frequency Presence
It is the method used, verified, and published by more authors [1, 2, 4, 5, 7, 10, 14]. The dependency is simple. If the fundamental frequency is detected in the voice segment, the segment is voiced, else the segment is surd.

The presence of the fundamental frequency can be verified by more methods of signal processing. It can be found as a peek in the real cepstrum of the signal [1, 2, 11]. The calculation of the cepstrum is slow. It can be substituted by faster methods, i.e. the autocorrelation [1, 2, 6, 11, 14-18].

2.1.2 Comparison of the Energy Spread
There is three or more frequency ranges defined, and the spread of energy in the all of these ranges leads to proper segment type determination. Each segment type has a typical spread of energy in the given frequency sub-ranges.

The preprocessing of the typical spreads is needed before use of this method; it leads to more time consumption [1, 2, 7, 10].
2.1.3 ZCR to Short-Time Energy Relation
Other method uses a relation of the mean value of the zero-crossing rate to short-time energy of the voice signal segment. The voiced segments have higher value of the short-time energy, and lower mean value of the zero-crossing rate. Both of the characteristics have relative values, and are defined without units. [1, 2, 3, 10]

2.1.4 Statistical Methods
Another methods use a statistical processing as well [8, 9, 14, 15]. The autocorrelation function applied on the signal segment provides an option to detect the presence of fundamental frequency and to estimate its value. [2, 3]

2.2 Phoneme Recognizer
Phoneme recognizer based on long temporal context was developed at Brno University of Technology, Faculty of Information Technology [12, 13]. It was successfully applied to tasks including language identification, indexing and search of audio records, and keyword spotting. Outputs from this phoneme recognizer can be used as a baseline for subsequent processing.

It is the open source tool. Source codes and binaries can be redistributed and/or modified under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or any later version. Model files can be used for research and educational purposes [12]. There are model files for the Czech, Hungarian, Russian, and English languages. The tool is able to recognize the phonemes from the audio signal (human voice, given language), and provides text output. The output consists of the information about the recognized phonemes, and its position in the processed sample of the voice signal. It is usable for the extraction of the particular phonemes from the given signal.

3 Experimental Verification
The voice segment type determination experiment was processed by two methods, and both result sets was compared and evaluated. There are six files of cardinal numerals of the Czech language recorded and used. These words are sound different each other. It is usable for the sufficient variability of the signal – see [11] for more details and additional information.

3.1 Process of Verification
The Fig.1 shows the process of verification. Each voice sample has the same duration, and the signal is divided to 70 segments. The voice segment type determination is processed by autocorrelation and by cepstral method.

The same sample of signal is processed by Phoneme Recognizer [12] to recognize phonemes of the word and its position in the given sample. Each phoneme is mapped to segments to be the segment type determination verified.

3.2 Results of Verification
Results of both determination method and the phoneme recognition are verified by mapping (see Fig.2 to 7 for all the used words). The first rows in these tables mean the number of the segment N. Next two rows show the voiced segments (marked as “X”) determined by autocorrelation A and cepstral method C.

![Fig.1 – Schema of the Verification Process]
In the last row, the position of the recognized phonemes $P$ is mapped. The “$<\text{sil}>$” symbol means the silence was recognized in the position. The Czech language character “ě” is phonetically identical with phoneme couple “je” in many words; it is corrected in the verification tables.

The verification denotes the results of the segment type determination and phoneme recognition match at first sight. The voiced segments are determined significantly in the position of vowels. It conforms to the expectation as is written above.

In the auxiliary view, some segments (i.e. Fig. 6, segment no. 62) were determined as voiced ones, but the phoneme recognition denotes, there was silence. It means the recorded voice signal had a higher noise level in that time. However, the rate of false determined segments is low by this type of error, because there are four cases only in the 420 processed segments.
For the recap of the previous experiments [11] – comparison of the voice segment type determination methods shows the difference in units of segments. The counter value of error rate is about less than 10% (see Tab.1 cited from [11]). It means, the difference between both methods can be omitted for the verification using phoneme recognition. Next work is to analyze the verification results.

<table>
<thead>
<tr>
<th>Compared word</th>
<th>Total segments</th>
<th>Autocorrelation</th>
<th>Cepstral method</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Voiced</td>
<td>Surd</td>
<td>Voiced</td>
</tr>
<tr>
<td>1  jedna</td>
<td>70</td>
<td>18</td>
<td>52</td>
<td>17</td>
</tr>
<tr>
<td>2  dvě</td>
<td>70</td>
<td>24</td>
<td>46</td>
<td>21</td>
</tr>
<tr>
<td>3  ří</td>
<td>70</td>
<td>19</td>
<td>51</td>
<td>17</td>
</tr>
<tr>
<td>4  čtyří</td>
<td>70</td>
<td>27</td>
<td>43</td>
<td>27</td>
</tr>
<tr>
<td>5  pět</td>
<td>70</td>
<td>10</td>
<td>60</td>
<td>7</td>
</tr>
<tr>
<td>6  šest</td>
<td>70</td>
<td>11</td>
<td>59</td>
<td>9</td>
</tr>
</tbody>
</table>

Tab.1 – Comparison Results [11]

The verification results analyses are provided in Tab.2 and Tab.3. There are noted counts of false determinations of the voiced type segment in the signal part recognized as the silence and as the surd phoneme. The absolute and corresponding relative values are denoted in the tables as well as the total values. The different results are highlighted in both tables to be well telling.

<table>
<thead>
<tr>
<th>Verified word</th>
<th>Total segments</th>
<th>False voiced type determ. out of the voiced signal ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  jedna</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>2  dvě</td>
<td>70</td>
<td>2</td>
</tr>
<tr>
<td>3  ří</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>4  čtyří</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>5  pět</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>6  šest</td>
<td>70</td>
<td>0</td>
</tr>
</tbody>
</table>

Tab.2 – Verification Results (Autocorrelation)

<table>
<thead>
<tr>
<th>Verified word</th>
<th>Total segments</th>
<th>False voiced type determ. out of the voiced signal ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  jedna</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>2  dvě</td>
<td>70</td>
<td>2</td>
</tr>
<tr>
<td>3  ří</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>4  čtyří</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>5  pět</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>6  šest</td>
<td>70</td>
<td>0</td>
</tr>
</tbody>
</table>

Tab.3 – Verification Results (Cepstral method)
4 Conclusion and Future Work

The results of the verification show that the counter value of false determination rate given by the voiced segment determination is under the error rate of both used determination methods [11]. It means the determination of the voice segment type using both methods corresponds to the phoneme recognition with the error rate fewer than 10%, and methods can be used very well.

This result will be used for the support of the theoretical base for the information system’s user identification by speaker’s voice. The next steps will be in the area of sufficient set of voice features extraction from the speaker’s speech. There are several methods that provide feature extraction from both types of the voice segment (voiced and surd). Therefore, the determination is the important part of preprocess for the extraction.

This type of user identification is simple for users. It is not necessary to remember passwords. The users can use standard multimedia inputs for the identification, none expensive equipment needed. The communication using multimedia is very popular nowadays. Multimedia applications as interactive systems of digital media are favored by many users [19]. The multimedia information can be selected by the user himself according to his individual needs, and the system can recognize the user by his/her voice. It can be the goal of this method.

5 Acknowledgement

This work was supported by the project No. CZ.1.07/2.2.00/28.0327 Innovation and support of doctoral study program (INDOP), financed from EU and Czech Republic funds.

References:
[12] P. Schwarz, P. Matejka, L. Burget, O. Glembek, Description Phoneme recognizer based on long temporal context, Brno 2012, online:


