Prosodic Cues for Automatic Word Boundary Detection in ASR

Klara VICSI, György SZASZÁK
Budapest University of Technology and Economics
Department for Telecommunication and Mediainformatics

Abstract: This article presents a cross-lingual study for agglutinative, fixed stressed languages, like Hungarian and Finnish, about the segmentation of continuous speech on word level by examination of supra-segmental parameters. We have developed different algorithms based either on a rule based or a data-driven approach. The best results were obtained by data-driven algorithms (HMM-based methods) using the time series of fundamental frequency and energy together. This HMM based method will be described in this article.

Word boundaries were marked with acceptable accuracy, even if we were unable to find all of them. On the base of this study a word level segmentation has been developed which can indicate the word boundaries with acceptable precision for both languages.

The evaluated method is easily adaptable to other fixed-stress languages.

Keywords: Automatic Speech Recognizers (ASR), word and phrase boundaries detection, prosodic recognizer, data-driven (HMM) approach, Finish language, Hungarian language

Introduction

Supra-segmental features are an integral part of every spoken language utterance. These can provide cues to the linguistic structure of the speaker’s message, emotional state or communicative intent. These features are intonation, stress, rhythm, etc. They can signal the structuring of utterances into larger discourse segments and provide additional information for human speech processing.

Using supra-segmental features in automatic speech recognition to increase its robustness is a tendency again [3,4,8,7]. Some trials were conducted in the mid-eighties, but it has not yet been possible to exploit such knowledge in an automatic speech recognition system [1,2,6]. This relative failure, according to Philippe Langlais [2], is mainly due to three types of difficulties:

- significant contextual variability of prosodic knowledge (type of speech, speaker, structure and content of sentences, nature of the environment, etc);
- complexity of relations between prosodic information and various linguistic organization levels of a message;
- problems encountered with accurate measurement of prosodic parameters, and their possible integration on a perceptual level.

---

1 Automatic Speech Recognition
Speech production is a continuous movement of the articulating organs, producing a continuous acoustic signal. Linguistic content and phonological rules help the human speech processing to separate syntactic units, such as words, phrases (sections between two intakes of breath), or sentences. In our experiments we have examined how words can be automatically separated in continuous speech in fixed stress languages such as Hungarian and Finnish (both of these languages belong to the Finno-Ugrian language family). These languages are highly agglutinative, so they are characterized with longer average word length than English, for example, and also with a relatively free word order. Stress is generally on the first syllable of the word, except in the case of conjunctions or articles. This means that the word-level intonation units we have dealt with during our experiment are composed of a word, stressed on the first syllable, together with unstressed conjunctions or articles, if any exist. This word level intonation unit is shorter than the phrases. Henceforward in this article we will strictly use the expression “word unit”, and the boundaries of these units will be called “word boundaries”.

A word level segmentationer has been developed capable to indicate the word boundaries with acceptable precision for both languages. The ultimate goal is to increase the robustness of speech recognition on the language modeling level by the detection of word boundaries, thus decreasing significantly the searching space during the decoding process. Searching space reduction is highly important in the case of agglutinative languages.

1. Methodology

In our experiment we have measured the fundamental frequency, the energy level and the time course. These parameters (all or some of them) are necessary for the realisation and perception of stress in the Finno-Ugrian language family. The peaks of energy and fundamental frequency clearly represent the first syllables of the words. Syllable length, it was found, is not greatly influenced by the stress.

If stressed syllables can be detected, the word boundary can be marked by finding the minimum of the energy and the fundamental frequency just before the stressed point. Hence the stressed vowels may be situated in the onset or even in the nucleus of the syllable, by searching the minimum point, the absolute minimum within an interval just before the stress should be found. The length of this interval is typically a half-syllable to avoid overlap with previous word. This value was set in empiric way, 100 ms interval length was found to be suitable for normal Hungarian read speech rate. We have developed different algorithms based either on a rule based or a data-driven approach [11]. The best results were obtained by data-driven algorithm for the determination of word units and their boundaries, therefore this approach will be described in this article.

Hungarian BABEL [5] and Finnish [10] continuous read speech databases were used for the examination. The databases were labelled and segmented on word level by an expert using audio visual method. For Hungarian 1600 sentences of 32 speakers and for Finnish 250 sentences of 4 speakers were used.
1.1. Acoustic Pre-processing

Fundamental frequency (Hz) and energy level (dB) were measured and their first and second order deltas. Syllable duration was found to be problematic to measure exactly without using speech recognizer output, hence it was discarded. For the detection of fundamental frequency, autocorrelation method was used. The autocorrelation function for x(n) discrete signal is:

\[ R(k) = \sum_{n=1}^{N-k} x(n)x(n+k) \]  

(1)

\( F_0(i) \) at the \( i^{th} \) frame \( F_0(i) \) was obtained after median filtering:

\[ F_0(i) = \text{med} \{ F_0(i-3), F_0(i-2), F_0(i-1), F_0(i), F_0(i+1), F_0(i+2), F_0(i+3) \} \]  

(2)

The energy \( E(i) \) was calculated with an integration time of 100 ms:

\[ E(i) = \frac{1}{M+1} \sum_{n=i-M}^{i} x^2(n) \]  

(3)

where \( M \) is the number of samples pro 100 ms.

The fundamental frequencies and the energy levels are given in every 25.6 ms.

1.2. Data-driven Approach

In data-driven approach the common HMM method [9] was used for the determination of word units and their boundaries. This type of examination needs a speech database segmented on word level to train the prosodic pattern HMM models. The word boundaries and the type of word units were marked in the speech database by an expert relying on fundamental frequency and energy cues, using audio-visual segmentation method. The segment boundaries of the word level prosodic curves were marked so that they overlap with word boundaries. 5 different word level prosodic curves, its boundaries and the silent periods were marked. Thus by the training a set of 6 different Hidden Markov Models were constructed. These models represent 5 types of word level prosodic curves which are descending, rising, floating, jumping and rise-fall while the 6th model is a silence model. Training examples for different curves are presented in Figure 1 between cursors. The sentence in Figure 1 was selected from the database in such a manner that it presents all of the used models in one example. However in the database 80% was descending, 10 % was silent and another 10 % was rising, floating, jumping and rise-fall together of all word units.
The prosodic model itself may be of interest in the future, but for the moment, we use only the boundaries of these word units for the evaluation.

We have also adapted this data-driven approach for Finnish language using the same 6 prosodic models as for Hungarian. Training strategy for Finnish was the same as for Hungarian. A cross- and multilingual experiment with Hungarian and Finnish was performed. In this case, we used prosodic HMM models trained on one language to segment speech on the other language. Moreover a hybrid was also trained to segment Hungarian and Finish speech.

1.3. Evaluation

To evaluate results the obtained prosodic segmentation is compared with the original one. Two measures are used to present our results. First one is the **correctness** which denotes whether a unit boundary predicted by our algorithm was correct or not (in %).

\[
\text{Corr} \ [\%] = \frac{\# \text{ of correctly marked word boundaries}}{\# \text{ of all marked word boundaries}} \times 100
\]

Our second measure is the **effectiveness** which says how many word boundaries were found at all (in %):

\[
\text{Eff} \ [%] = \frac{\# \text{ of correctly marked word boundaries}}{\# \text{ of all real word boundaries (in reference transcription)}} \times 100
\]

Note that this second measure is expected to be less than 100 %, since not all of the words in speech are emphasized. In this way articles and conjunctions can never be separated because they are not emphasized.

Since for speech recognition tasks we require that whenever a word boundary is predicted it should be detected correctly the correctness is more critical than the effectiveness. (We have required 80% correctness at least.) Of course, the higher effectiveness the more robust the system, but we cannot permit this at the expense of lower correctness.

The predicted segment boundary of word units is regarded as correct if it deviates no more than 2 frames forward or backward from the real word boundaries.
1.4. Optimization of main HMM parameters

By training of the HMM prosodic recognizer the first task was the optimization of the main HMM parameters. This means mainly to find the optimal number of states: since word level prosodic units are longer than phonemes – modeled usually by 5 state models at 10 ms frame rate – the number of states for prosodic models can be changed between 9 and 15. We have applied a frame rate of 25.6 ms so a 9 state model requires a minimal phrase length of ~230 ms while minimal phrase length for a 15 state model is ~380 ms which is according to our examinations a typical value for Hungarian. The optimal number of states was found to be 11 as it can be seen on Figure 2.

Only 6 dimensional observation sequences (F0, energy, first and second order deltas) were used and 2 or at most 4 Gaussian components were sufficed to describe output distribution of each state. The first and the last states were non-emitting ones [9]. Two acoustic pre-processing alternatives were examined with the above mentioned fixed HMM parameters: In the first case we have used either F0 or Energy data only with appended first and second order deltas while in the second case both F0 and Energy data have been applied as input with their first and second order deltas appended.

Table 1. Correctness and effectiveness of word unit’s boundary determination with different acoustical preprocessing

<table>
<thead>
<tr>
<th>Used prosodic parameter(s)</th>
<th>Language</th>
<th>Training corpus</th>
<th>Correctness/Effectiveness [% / %]</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0+ΔF0+Δ²F0</td>
<td>Hungarian</td>
<td>14 persons</td>
<td>67.4 / 58.4</td>
</tr>
<tr>
<td>E+ΔE+Δ²E</td>
<td></td>
<td></td>
<td>67.7 / 66.6</td>
</tr>
<tr>
<td>(F0+ΔF0+Δ²F0 + (E+ΔE+Δ²E)</td>
<td></td>
<td></td>
<td>76.5 / 53.0</td>
</tr>
</tbody>
</table>
In Table 1 the results are presented as a function of the acoustic pre-processing parameters. The best results were obtained when fundamental frequency-type parameters were used together with energy-type parameters. We have also tested several training strategies for the constructed HMMs. During the examination a 14 speaker data set (14x50 sentences) was used for training and 18 speakers (18x50 sentences) for testing. First the size of the training material was changed. The training set was reduced to 4 persons (4x50 sentences) and finally to one person (50 sentences), while the test set consisted of the same 18 speakers (18x50 sentences) in all cases.

Table 2 Correctness and effectiveness of boundary determination of word unit with HMM for different training settings.

<table>
<thead>
<tr>
<th>Used prosodic parameter(s)</th>
<th>Language</th>
<th>Training corpus</th>
<th>Correctness/Effectiveness [% / %]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(F0+ΔF0+Δ2F0+(E+ΔE+Δ2E))</td>
<td>Hungarian</td>
<td>1 person</td>
<td>77.1 / 49.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 persons</td>
<td>77.4 / 57.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 persons</td>
<td>76.5 / 53.0</td>
</tr>
</tbody>
</table>

If the HMMs were trained on few speakers these speakers were carefully selected in order to ensure a relatively accurate training corpus. Results are shown in Table 2. Surprisingly, there is no relevant difference in correctness if fewer speakers are involved in the training corpus. Effectiveness, however, depends very much on the number of speakers, and to achieve effectiveness over 50% at least four speakers’ data should be used for training. In comparison with the results obtained by the rule-based approach effectiveness is much higher with an acceptable correctness. If we use only $F_0$ or energy patterns, effectiveness is excellent, but we have ~10% reduction in correctness. The overall best result was 77.4% correctness with 57.2% effectiveness, obtained with HMMs trained on 4 speakers’ $F_0$ and energy data.

For further experiments the optimized parameters were applied: frequency and energy level values together, the speech examples of 4 persons and the state of the models were 11.

2. Results

The developed system can be convenient for automatic segmentation of word units. An example is presented in Figure 3 of how the developed segmentation method works on word level. Time function of the speech signal, $F_0$ and energy level contour are visible on the screen. The first row at the bottom contains an expert-made hand segmentation taken as reference while the second row illustrates the output obtained by the automatic prosodic segmentation. Segmentation accuracy means the correctness of the word unit boundaries’ determination.
Figure 3 HMM provided word level segmentation versus expert-made hand segmentation on a passage of 3 Hungarian sentences: [s] and [k] denote word boundaries, [sil] silence

Results obtained for Finnish with the same data-driven method as for Hungarian are presented in Table 3. We may note a small fall in correctness while effectiveness is high. The reason for this may be that Finnish speech is much slower than Hungarian, and Finnish words often contain long plosive sounds where $F_0$ and energy contour show a very similar behavior to the one they have on real word boundaries. As a result, many more word boundaries can be found than in

<table>
<thead>
<tr>
<th>Used prosodic parameter(s)</th>
<th>Language</th>
<th>Training corpus</th>
<th>Correctness/Effectiveness [% / %]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_0 + \Delta F_0 + \Delta^2 F_0$ + $E + \Delta E + \Delta^2 E$</td>
<td>Hungarian</td>
<td>4 persons</td>
<td>77.4 / 57.2</td>
</tr>
<tr>
<td></td>
<td>Finnish</td>
<td>4 persons</td>
<td>69.2 / 76.8</td>
</tr>
</tbody>
</table>

Hungarian, but we can also detect some in-word secondary emphasis which explains the lower correctness.

Cross-lingual prosodic word boundary segmentation results both for Finnish and Hungarian are presented in Table 4. It can be seen that segmentation of Finnish speech with models trained on a Hungarian database gives nearly the same result as the

<table>
<thead>
<tr>
<th>HMM models</th>
<th>Test data</th>
<th>Corr/Eff [%/%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hungarian (4 speakers)</td>
<td>Hungarian</td>
<td>77.4 / 57.2</td>
</tr>
<tr>
<td>Hungarian (4 speakers)</td>
<td>Finnish</td>
<td>67.1 / 52.1</td>
</tr>
<tr>
<td>Finnish (4 speakers)</td>
<td>Finnish</td>
<td>69.2 / 76.8</td>
</tr>
<tr>
<td>Finnish (4 speakers) + Hungarian(4 speakers)</td>
<td>Hungarian</td>
<td>70.7 / 52.3</td>
</tr>
<tr>
<td>Finnish (4 speakers) + Hungarian(4 speakers)</td>
<td>Finnish</td>
<td>75.0 / 68.2</td>
</tr>
<tr>
<td>Finnish (4 speakers) + Hungarian(4 speakers)</td>
<td>Finish</td>
<td>69.7 / 83.7</td>
</tr>
</tbody>
</table>
Finnish models. On the other hand, segmented Hungarian data with Finnish models yields 70.7% correctness which is somewhat better than on Finnish data (67.1%). This is probably due to the sparseness of Finnish data.

Generally this means that a prosodic word boundary segmentator well trained with a Hungarian database can be applied for automatic segmentation of unknown Finnish speech material and vice versa. Naturally hand-made correction is necessary.

3. Conclusion

Our word level prosodic segmentation method based on measuring fundamental frequency and energy level functions gave promising results. Word boundaries can be marked with acceptable correctness, even if we are not able to find all of them. Two measurements, correctness and effectiveness, were used to describe the behavior of this prosodic segmentation system.

Word boundaries are found with acceptable correctness and effectiveness for fixed-stress languages like Hungarian and Finnish. In case of Finnish we obtained results comparable with Hungarian with lower correctness and higher effectiveness which may be the result of the difference between the two languages and also of data sparseness in the Finnish database.

Moreover, these results ensure that the integration of a prosodic recognizer into a CSR system can help reduce the searching space and thus improve speech recognition performance. The importance of this searching space reduction is great in recognition of agglutinative languages such as Hungarian and Finnish where the possible number of word forms is more than hundreds of millions. In this domain further investigations are needed.

Summarizing the results of our experiments it is clearly worth continuing research in this field. The evaluated method is easily adaptable to fixed-stress languages other than Finnish and Hungarian. The examination of such languages would be useful.

We have presented a single example how it is possible to use prosodic information, but there may be several solutions.

A practical result emerged from these experiments: this prosodic recognizer can be used as a word-level automatic segmentationer for Hungarian and Finnish languages.

4. Acknowledgement

We would like to express our thanks to Toomas Altosaar (Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing) for his kind help and his contribution to the use of the Finnish speech database. The work has been supported by the Hungarian Scientific Research Foundations OTKA T 046487 ELE and IKTA 00056.

References


