Pronunciation Modeling in Hungarian Number Recognition

Tibor Fegyó, Péter Mihajlik, Péter Tatai, and Géza Gordos

TSP Lab., Department of Telecommunications and Telematics
Budapest University of Technology and Economics, Hungary
fegyo@tt.ttt.bme.hu

Abstract
In Hungarian, as more or less in many other languages, a large percent of words and phrases can be pronounced in several, different, but correct ways. Introducing pronunciation alternatives for individual vocabulary elements may improve the efficiency of the recognition. But in connected word recognition tasks the modeling of inter-word phonetic changes has a greater significance. In this paper we introduce a rule-based method for the automatic generation of pronunciation alternatives used first for isolated words and later the method is extended to handle cross-word phonological changes in recognition networks, applying a special approach applicable for the Hungarian language. To evaluate the method it is tested in connected number recognition tests.

1. Introduction
In the last decade various pronunciation modeling techniques have been developed to bridge the gap between lexicon and pronounced speech [1]–[3]. Though notable results have been achieved, the studies concentrated mainly on isolating languages, like English. For these languages it is sufficient to model the pronunciation alternatives inside of the words and at word boundaries as specified by the N-gram language model.

For strongly agglutinating, morpheme-based languages, like Hungarian, this approach, however, is not directly applicable. In this study we first present our experiments in the field of isolated word recognition. The significance of the pronunciation alternatives inside of the words is investigated, compared to the case when only one (the most probable) transcription per word is allowed. For the experiments a rule-based tool developed at our Laboratory has been used to generate automatically the phonetic transcriptions. This task is intended to serve as a baseline for further investigations.

The main goal of our examinations is to determine how much improvement (if any) can be achieved by applying the same phonological rules as used in isolated recognition, but this time on a morpheme-network used for connected or continuous recognition. As for Hungarian and other agglutinating languages the N-gram language models are not suitable, a morpheme-based grammar network should be used. In such networks the adequate modeling of phonological interactions at the morpheme boundaries is essential, and the known techniques for stochastic language models cannot be applied.

To undertake this task we needed such a recognition problem, where the possible variations of word (morpheme) connections are all known so they can be represented in a treatable size grammar network. Therefore, we have chosen the morpheme-based connected word number recognition task. In the experiments we investigated different approaches for inter-morpheme pronunciation modeling and their effect on the recognition error rate.

Although our experiments relate to a special problem – number recognition in Hungarian, a relatively small language spoken by about 15 million people – the results might be extrapolated for other situations and other languages, too, with some care, to yield valid knowledge for the speech research community.

2. Automatic Phonetic Transcription for Hungarian
In this section we very briefly launch the problems related to the automatic phonetic transcription of Hungarian, introduce a method which solves these problems for isolated words, and explain why they are not directly applicable to connected or continuous recognition.

2.1. Automatic Pronunciation Modeling of Isolated Hungarian Words
In Hungarian, there is a relatively close correspondence between orthography and pronunciation, so automatic transcription is considered as a relatively easy task. But this is valid only for the canonical form. Unlike in English, in Hungarian it is not generally true that a word is pronounced most probably according to its canonical transcription. In order to generate the actually pronounced phonotypical forms, the phonological interactions (assimilations, etc.) between neighboring phonemes must be taken into account because they occur very frequently. In most cases these phonetic interactions appear at the boundaries of adjacent morphemes.

Due to the allowed variations in pronunciation the phonological changes usually cannot be predicted in advance in every particular case, e.g. “d’ s” may become “d’ s” or “t’ s” or “ts’”. (For speech sounds the SAMPA representation is used. [6]) Thus, the phonotypical transcription should naturally include the pronunciation alternatives.

In the following an algorithm developed at our Laboratory is presented, which is applicable to the automatic generation of phonotypical transcriptions of Hungarian words by means of phonological rules and morpheme analysis.

The main steps of the algorithm (see details in [4]):
1. Morphemic segmentation of the orthographic form of the word. Example: egyszér → egyszzer (once) (“z” and “sz” represent special morpheme boundary information.)
2. Segmentation to graphemes on a per morpheme basis. Example: egyszer
3. Grapheme – phoneme conversion: egyszer
   Example: ed’ s e r (canonic form)
4. Application of phonological rules: 
Example: e <d'> s | t' s | ts: > e r (phonotypical form)

5. Pronunciation network generation.

![Diagram of pronunciation network]

Figure 1: The automatically generated pronunciation network of the Hungarian word *egyszer* (once) with SAMPA symbols.

Finally, the graph representations of the vocabulary elements can easily be connected in parallel with each other in the recognition network.

2.2. The Need for Cross-Word (Inter-Morpheme) Pronunciation Modeling

If the pronunciation network of the words are connected after each other and the previously outlined method is used, the phonetic changes at the word boundaries will not be represented in the recognition graph.

One may propose to join all the commonly used Hungarian words in parallel to estimate the N-gram language model, and to model the maximum P X P cross-word pronunciations (where P is the number of phonemes in Hungarian), and to solve in this way the LVCSR (Large Vocabulary Continuous Speech Recognition) problem in Hungarian. Unfortunately, this is not feasible, because the most commonly used forms of Hungarian words are well above several millions due to the agglutinative nature of the language. In fact, the basic building units of our native language are not the words, but the morphemes of which there are less then one hundred thousand.

The order of the morphemes within one word can formally be described. The generally used N-gram language model estimation may result in a less accurate model [4]. So the morpheme-network vocabulary representation (including the knowledge based language model) seems to be the best way towards the LVCSR of Hungarian, but for that aim, the cross-word and inter-morpheme pronunciation modeling must be solved, too.

The method described in the previous subsection can be extended for handling this problem in the following way:

(i) perform the first three steps of the algorithm for all the vocabulary elements (morphemes or words), and then

(ii) convert them to graph format, connect with each other as required, so resulting in a *canonic morpheme-network*.

Finally,

(iii) apply the phonological rules to this network, to yield the phonotypical recognition graph, which includes the pronunciation alternatives, both inside and at morpheme boundaries.

Obviously, the implementation of the final step for general-structure networks is not straightforward. To establish an (in some sense) optimal solution for this, we chose a partial problem for experimentation, namely, the Hungarian number recognition task.

3. Recognition of Hungarian Numbers

3.1. Isolated Word Based Recognition

To analyze the efficiency of the phonetic transcriber algorithm and to have a basis of comparison for the results of continuous number recognition tests, isolated word based number recognition is investigated. The vocabulary itself contains all the possible numbers with their automatically generated phonetic transcription. The Hungarian numbers are composed as connected words (e.g., 621 = *hat* + *száz* + *huszon* + *egy*). As mentioned, on word and morpheme boundaries there are sometimes pronunciation alternatives. For example, in the previous example the “i” and “s” sound can remain two separated sound or can melt into a “tsi”. The phonetic transcription procedure can alternatively list only the most frequent form, or all of the possible forms to test if they are real alternatives. Our recognition experiments proved that the alternatives really occur in the spoken language.

3.2. Connected Word Based Number Recognition

For large recognition tasks it is not efficient to list all of the words in the vocabulary. If the task were to recognize the numbers from 0 up to 999,999 then, as many numbers have multiple forms, more than 2 million items should be included in the vocabulary. The general solution could be to build a network to generate all possible forms. This network was constructed for the Hungarian language.

Each node represents a number building item like *egy*(one), *tiz*(ten), *tizen*(...teen), *ezert*(thousand)... In the vocabulary the pronunciation transcription of these units is declared. As before, the transcription is automatic, but in this case the cross-word changes need special considerations.

For example, the unit *száz*(hundred) has multiple pronunciations depending on the context. As a standalone number it is pronounced as “s a: z”. If the preceding number ends on a phoneme “í” then they melt into a “tsi” or may also be pronounced as separated sounds. On the other hand if the subsequent number begins with an unvoiced sound, then the ending “z” sound gets unvoiced (“s”). So, as it has been illustrated, such units may have multiple pronunciations depending on the context.

3.3. Possible Language Models

To handle the above phenomenon we developed a method, which extends the general canonic language model to a phonetically correct network. In the following this will be described in details.

3.3.1. Basic Network

![Diagram of basic network]

Figure 2: Part of the basic network

The basic network is constructed in accordance with the general solution. For our purpose a network is used with a vocabulary listing, one pronunciation for each unit. In this...
case the cross-word phonological changes are not considered, because not all of them is correct, and there are missing pronunciation forms, too. In Figure 2 the word száz should have two distinct forms according to the following items: száz + egy is spelled "s a: z e d' ", but száz + húsz is spelled "s a: s h u: s," but "s a: z h u: s" is incorrect.

3.3.2. Extended Vocabulary

An obvious solution to handle alternative pronunciations is to extend the vocabulary with all the possible pronunciation alternatives. The alternatives are generated from word pairs occurring in the task. So the word száz has four different forms as the initial sound can be either "s" or "ts" and the ending sound can be either "s" or "z", but word egy cannot be followed by any unvoiced sound; so the ending d' has no alternative. The baseline network is not changed, only the vocabulary is extended (Figure 3). The disadvantage of this solution is that the model will generate many inaccurate forms e.g., " s a: s e d' ". (In Figure 3 the dashed line represents the incorrect links.) Based on our recognition experiments, however, in the following sections the usefulness to enable multiple pronunciation alternatives is demonstrated.

![Figure 3: Part of the basic network using the extended vocabulary](image)

3.3.3. Extended Network

All erroneous pronunciation forms should be elimated by the recognizer. An equivalent solution can be obtained with the extended vocabulary by allowing multiple pronunciations at the network level as parallel transitions. In this case a bigram language model would score down the incorrect junctions. However, it should be trained with a very large set of data and smoothed, furthermore, the estimation is not always accurate. In the Hungarian language – as mentioned before – the cross-word pronunciation changes have well defined rules. These rules are formalized, so they can be applied to transform a canonical network into a phonetically correct representation. This process is not fully automatized yet, but a network was built for the number recognition task and adjusted partially by hand. According to the experiments on number recognition, this method overperformed all previous solutions.

The first step in creating the extended network is to build a phoneme-based network: the items in the graph are substituted with the corresponding phoneme sequence from the vocabulary. The inserted phonemes are placed in the nodes of the graph. In the second step each node is duplicated if the corresponding phoneme must be changed because of the type of the preceding or the following phoneme and in the new node the corresponding new sound is inserted. Finally, the links connected to the original node are revised and moved (or duplicated) to the new node, if needed, according to the applied rules.

![Figure 4: Part of the optimized network](image)

In Figure 4 the first and last phoneme of száz is duplicated. The “s” can melt with the preceding “t” to a “ts:” sound and the “z” gets unvoiced due to the following “h". After duplicating the “s:” node, only the correct links are kept.

4. Experimental Results

A set of experiments was carried out in order to assess the relevance of modeling to the phonetic changes at word boundaries, and to the use of pronunciation variants in number recognition. Both isolated and connected word based tests were made.

4.1. Model Training

Currently BABEL is the largest high quality Hungarian speech database available for research purposes [5]. It consists of three different parts: isolated digit and continuous number utterances, consonant-vowel-consonant syllables, and continuously read speech. A fraction of the database is segmented and labeled by hand at the phoneme level. We had access to the labeling information for five speakers’ voices, 400 seconds of speech in total. The number of speakers available is 20 (10 men and 10 women), and there are altogether about 900 sentences and 9700 continuous number strings in the database. The voice of 14 speakers composed the training set, and the rest of the data were used in the recognition tests. In the experiments the numbers and the paragraphs were used separately for training, resulting in 2 different acoustic model sets. For parameterization we used 10 MFCC + log energy with delta and acceleration coefficients and alternatively 16 LPC cepstra + log energy with delta coefficients.

Embedded Viterbi-training was applied with the FlexiVoice tool [4]. The models were left to right ones with three states per phoneme. A 10-mixture diagonal covariance matrix Gaussian distribution has been used in all states. From among the 64 phonemes of Hungarian, only the vowels and the short consonants were used [4].

4.2. Isolated Number Recognition

During the isolated number recognition tests all 140 numbers occurring in the test database are listed in the vocabulary. The numbers are transcribed to phoneme sequences automatically. In the experiment the effect of the presence or the absence of pronunciation alternatives were investigated (Table 1). In the first case the most frequent pronunciation was used while in the second case all alternatives were listed in the vocabulary.
In the second group of experiments the vocabulary contained 4,300 connected word based number recognition test was misrecognized, so the word error rate is even lower. Moreover, in most of the cases only one item in the number set confirmed the importance of pronunciation alternatives. The experiments were carried out using both training sets, and both parameterization methods (Table 2).

The error rate decreased slightly for both acoustic model sets confirming the importance of pronunciation alternatives. Moreover, in most of the cases only one item in the number was misrecognized, so the word error rate is even lower.

4.3. Connected Word Based Number Recognition Test

In the second group of experiments the vocabulary contained about 50 numbers (e.g., egy (1), egyes (1), tiz (10), tizen (10), tizes (10),...), and a network generated all possible numbers. The three different networks described in section 3.3 were investigated. The basic network generated the numbers from 0 to 999,999 and only intra-word phonetic changes were considered. In the extended vocabulary the base morphemes were the same, but all possible pronunciation alternatives were listed that occur in this task. In the last case the network was phoneme based, and it allowed only the proper pronunciation possibilities. The experiments were carried out using both training sets, and both parameterization methods (Table 2).

Table 1: Isolated number recognition error rates using two different pronunciation models. Acoustic models were trained by numbers (a) and by general speech (b) using MFCC parameters.

<table>
<thead>
<tr>
<th>Vocabulary representation</th>
<th>Error rate</th>
<th>Relative improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most frequently used form</td>
<td>0.48%</td>
<td></td>
</tr>
<tr>
<td>Pronunciation alternatives</td>
<td>0.45%</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

Table 2: Connected word based number recognition error rates using various pronunciation models. Acoustic models were trained by numbers (a) by general speech (b) using MFCC parameters; and by numbers (c) using LPC cepstra parameters.

<table>
<thead>
<tr>
<th>Network type</th>
<th>Digit string error rate</th>
<th>Relative improvement</th>
<th>Digit string error rate</th>
<th>Relative improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic network</td>
<td>5.89%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extended vocabulary</td>
<td>3.99%</td>
<td>32.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-morpheme phonetic changes with alternatives</td>
<td>3.45%</td>
<td>41.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic network</td>
<td>16.28%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extended vocabulary</td>
<td>13.42%</td>
<td>17.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-morpheme phonetic changes with alternatives</td>
<td>13.28%</td>
<td>18.4%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If cross-word changes were not modeled, the error rate became the highest, but modeling the phonetic changes improved the system. With a brute force solution the relative improvement was lower, while with the optimal network, a higher improvement could be achieved. The results indicated that the best acoustic models resulted from using the MFCC parameters and a training set containing only numbers. The highest relative change between the basic and the revised pronunciation models was achieved using the best acoustic models, but the improvement is significant in the other cases, too.

The error rates are larger even in the best case than they were in the isolated number tests, but it can be noted that the effective vocabulary size grew from about 200 up to more than 2,000,000.

5. Conclusion

In this article we presented a rule-based method, which is capable to produce automatically the various pronunciation forms of Hungarian words. Also, we introduced two inter-morpheme pronunciation modeling methods and applied them for the connected word based number recognition task. In the first case a basic vocabulary was extended to represent all the pronunciation alternatives of the morphemes. In the second, more elaborated modeling technique, an optimal phoneme based network was generated enabling exclusively the correct pronunciation forms.

Finally, isolated and connected word based number recognition test results were presented. The introduction of pronunciation alternatives themselves did not affect significantly the recognition in the isolated task. But the application of the phonological rules at morpheme boundaries resulted in a substantial improvement in connected recognition tasks. A noticeable phenomenon was that the higher the baseline error rate was, the smaller the relative improvement caused by the pronunciation modeling became, which means that the quality of the acoustic model deeply influences the effectiveness of the pronunciation modeling.

6. References