Investigation of Morph-based Speech Recognition Improvements across Speech Genres

Péter Mihajlik \(^1,2\), Balázs Tarján \(^1,3\), Zoltán Tüske \(^1\), and Tibor Fegyó \(^1,3\)

\(^1\)Dept. of Telecom. and Media Informatics, Budapest University of Tech. & Economics, Hungary
\(^2\)THINKTech Research Center, Hungary, \(^3\)AITIA International Inc., Hungary

\{mihajlik, tarjanb, tuske, fegyo\}@tmit.bme.hu

Abstract

The improvement achieved by changing the basis of speech recognition from words to morphs (various sub-word units) varies greatly across tasks and languages. We make an attempt to explore the source of this variability by the investigation of three LVCSR tasks corresponding to three speech genres of a highly agglutinative language. Novel, press conference and broadcast news transcription results are presented and compared to spontaneous speech recognition results in several experimental setups. A noticeable correlation is observed between an easily computable characteristic of various broadcast news transcription results are presented and based LVCSR results across four languages \([4]\). The difficulty is partitioned into 5 categories, including spontaneous speech, for press conference speech, and for classical broadcast news speech. All experiments are performed in Hungarian – as one of the languages with high morphological complexity – so that cross-lingual effects do not bias the comparison. Press conference and broadcast news transcription systems are completely novel achievements for Hungarian, not presented before. The conclusions are extended for other languages, as well.

2. Concept and Tasks

Since morph-based speech recognition results scatter most heavily on a speech genre scale – from read to spontaneous conversational speech – our concept is to measure the improvements due to morph-based speech recognition on a spontaneity scale. Three points on this scale corresponding to three Hungarian language LVCSR tasks are examined.

- **Spontaneous speech / MALACH task (SP):** probably the best choice would be to analyze a Callhome or Switchboard type Hungarian corpus at the spontaneous end of the scale. Unfortunately no such type of speech database is available in Hungarian, but we had access to the Hungarian MALACH data. This corpus contains interviews with elderly people and is detailed in \([6,7]\).
- **Mostly planned speech / Press Conference task (PC):** the press conference audiovisual material of the Hungarian government is publicly available. What makes this LVCSR task attractive is that all the transcriptions of press conferences are open for the public for years – altogether 1.2 million words. However, the transcriptions are not always exact, disfluencies and noises are not marked and ungrammatical sentences are corrected. Questions from press people and answers are included in the data, only unintelligible recordings are removed. The amount of test data is 4 hours, 19K words (matched data set in \([6]\)).
- **Planned speech / Broadcast News task (BC):** we used publicly available broadcast news audiovisual data of a Hungarian TV channel specialized for news. Unfortunately no transcriptions are available, but a relatively large amount of broadcast news text data is placed on the website of the channel (5.6 million words). The recordings consist of basically clean speech. 1 hour of speech corresponding to 7.7K words is used as test data in the experiments.
In the followings word- and morph-based speech recognition approaches and the results of the three LVCSR tasks are presented and analyzed. The aim is to explore the dependencies of the improvement due to morph-based speech recognition, i.e. which factors affect the improvement crucially and which are less critical. A further aim of the investigations is to understand the different behavior of spontaneous and planned speech results.

3. Methodology

The differences between morph- and word-based speech recognition results are measured in several experimental setups. In each setup both word- and the morph-based systems are built and optimized on the given, task specific “in domain” training text database as follows.

3.1. Text corpus preparation

Whereas no extraordinary corpus preparation is required for word-based speech recognition, morph-based systems need special treatment of the given training text data. In our approach, first word boundary symbols <w> are placed into the text, after each word, and are considered as separate morphs. (<w> symbols are required for the reconstruction of word boundaries in the decoder output [8]). Then a core word list is collected leaving out all special tokens like acronyms, abbreviations, etc. Morph segmentation is performed on this core word list resulting in a “word-to-morph” dictionary. The corpus for a morph-based speech recognition system is obtained by replacing each word of the corpus by the corresponding morph sequence. The words not presented in the word-to-morph dictionary remain intact in the corpus and treated as simple morph tokens in the subsequent operations.

3.2. Speech recognition models

3.2.1. Language model

In each setup, both word- and morph-based n-gram language models are built on the correspondingly processed task specific corpora with full vocabularies applying the modified, interpolated Kneser-Ney smoothing technique [9] implemented by the SRILM toolkit [10]. Depending on the task, on the type of morphs and on the training corpus size word and morph vocabulary sizes are in the range of 20k – 285k and 5k – 80k, respectively. By default, full 3-gram language models are built for the words and full 4-gram models for the morphs (ignoring 3 and 4 grams found only once). The only exception is at the 5.6M BC corpus, where entropy-based pruning [11] is applied both on the word and on the morph n-grams resulting in roughly equal language models sizes in term of number of grams.

3.2.2. Pronunciation model

Simple grapheme-to-phoneme rules [12] and exceptions are applied in order to obtain word- and morph-to-phoneme mappings. Weighted alternative pronunciations are used only for the SP (MALACH) task, though as [3,7] showed, their effect is minimal on the recognition accuracy.

While there is a virtual “os = optional silence” model at the end of each word’s pronunciation (with similar aim to the so-called “sp” model [13]), no such model is attached to the pronunciation of morph models. Instead, the <w> symbol itself is mapped to the “os” model.

3.2.3. Context dependency model

As equation (1) shows, triphone context expansion is performed after the integration of higher level knowledge sources, so that context dependency is modeled across word- and morph-boundaries, with respect to inter-word optional silences, as well.

3.2.4. Acoustic models

Speaker independent decision-tree state clustered cross-word triphone models were trained using ML (Maximum Likelihood) estimation [13]. Three state left-to-right HMMs were applied with GMM’s (Gaussian Mixture Models) associated to the states. For the SP task, 26 hours of “in domain” training speech was used for training 3000 HMM states with 10 Gaussian per state, based on PLP (Perceptual Linear Prediction) features [3]. For both the PC and BC tasks the acoustic models were trained on the MRBA database [14] augmented with about 10 hours of transcribed PC speech. In that case the number of states was about 2500 and 8 Gaussians were used per state. The feature type was MFCC (Mel-frequency Cepstral Coefficients) with delta and delta-delta, calculated on 8 kHz bandwidth speech and blind channel equalization [15] was also applied.

3.3. Off-line recognition network construction

The WFST (Weighted Finite State Transducer) [16] recognition network is computed on the triphone-level:

\[
\text{wred}\left(\text{fact}\left(\text{compact}(C \circ S \circ \text{compact}(\text{det}(L \circ G)))\right)\right)
\]

where capital letters stands for transducers, others for operators detailed below. First the language model (G) and pronunciation model (L) is composed and determinized. Then some auxiliary symbols are removed and a suboptimal minimization procedure – called as compaction – is applied that does not need the argument to be determinizable. Then each “os” model is replaced to a null-transition and to a normal silence model switched parallel by using a simple (S) transducer. The next step is the triphone context expansion (C), then the WFST network is compacted, factorized and the weights are redistributed resulting in a stochastic transducer suitable for the WFST decoder.

3.4. Evaluation

One-pass decoding was performed by the frame synchronous WFST decoder called as VOXserver – developed in our laboratories. RTF (Real Time Factor) of a morph- and the corresponding word-based system were adjusted to be close to equal using standard pruning techniques. RTF values were about 1 for small and midsize training text corpora for the PC and BC tasks, and about 4 for the largest corpora and for the SP task – measured on the same 3GHz CPU.

The SP and the BC test sets contain only the speech of previously unseen speakers. All the PC and BC test speeches arose later in time than the related training text data. Though in case of morphologically rich languages WER (Word Error Rate) – to some extent – shows a pessimistic picture of the speech recognition performance [4], we used it as the basis of evaluation since it is the most widely accepted and interpretable measure. Under the term of ‘improvement’ WER reduction is understood.

Signed-rank Wilcoxon tests with a significance level of 0.05 were applied to judge if a morph-based improvement is significant over the corresponding word-baseline.
4. Experimental results

First full-scale results are presented on the three Hungarian LVCSR tasks with various morph types. Then, we make attempts to converge the results of the three speech genres by using equal size training text corpora and with down- or “upgrading” acoustic conditions.

4.1. Full-scale results of various morph types

The following morph types – in term of the applied vocabulary decomposition algorithm – are evaluated on all speech genres exploiting full training text corpora:

- Statistical morphs: selected words are segmented to morphs by using the unsupervised MB (Morfessor Baseline, [17]) and MC (Morfessor Categories-MAP, [18]) algorithms.
- Grammatical morphs: morphs were obtained by applying an affix-stripping method implemented in the Hunmorph system [19,20]. Two methods are used, a grammatically strict HSF (Hunmorph Strict Fallback) and a less strict, more heuristic HCG (Hunmorph Compound Guessing).
- Combined morphs: CHM (Combined Hunmorph Morfessor) the MB algorithm is used to disambiguate the multiple morph analyses of Hunmorph system [6,7].

More detailed description of these morph types can be found in [3], [6] and [7].

Table 1. Full-scale results of various morph types.

<table>
<thead>
<tr>
<th>Task/meas.</th>
<th>Word</th>
<th>MB</th>
<th>MC</th>
<th>HSF</th>
<th>HCG</th>
<th>CHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP WER [%]</td>
<td>52.9</td>
<td>51.3</td>
<td>50.6</td>
<td>51.3</td>
<td>50.9</td>
<td>50.5</td>
</tr>
<tr>
<td>SP -Δrel [%]</td>
<td>-</td>
<td>3.0</td>
<td>4.4</td>
<td>3.0</td>
<td>3.8</td>
<td>4.5</td>
</tr>
<tr>
<td>PC WER [%]</td>
<td>43.5</td>
<td>34.5</td>
<td>34.5</td>
<td>34.9</td>
<td>35.6</td>
<td>35.2</td>
</tr>
<tr>
<td>PC -Δrel [%]</td>
<td>-</td>
<td>21</td>
<td>21</td>
<td>20</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>BC WER [%]</td>
<td>36.7</td>
<td>21.0</td>
<td>21.4</td>
<td>23.1</td>
<td>24.2</td>
<td>20.8</td>
</tr>
<tr>
<td>BC -Δrel [%]</td>
<td>-</td>
<td>43</td>
<td>42</td>
<td>37</td>
<td>34</td>
<td>43</td>
</tr>
</tbody>
</table>

As the results in Table 1 show, there are huge differences between the relative improvements across LVCSR tasks. All the error rate reductions are significant as well, however, with only two exceptions, no significant differences can be observed between the performances of the various morph-based results in a given task. Therefore only the MB-type morphs – where word-to-morph segmentation is obtained by the Morfessor Baseline method – are investigated further.

4.2. Effects of down-scaled training text corpora

In order to eliminate the effect of differently sized training texts, we made additional experiments based on equal-size training text corpora. The most recent texts are left in the reduced training databases of the PC and BC tasks.

Table 2. Word- and morph-based results with full and with down-scaled training text corpora.

<table>
<thead>
<tr>
<th>Task</th>
<th># of training words</th>
<th># of word forms</th>
<th>OOV rate [%]</th>
<th>Word WER [%]</th>
<th>MB WER [%]</th>
<th>-Δrel [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td>160k</td>
<td>20k</td>
<td>15.6</td>
<td>52.9</td>
<td>51.3</td>
<td>3</td>
</tr>
<tr>
<td>PC</td>
<td>160k</td>
<td>26k</td>
<td>14.1</td>
<td>53.1</td>
<td>43.0</td>
<td>19</td>
</tr>
<tr>
<td>BC</td>
<td>160k</td>
<td>30k</td>
<td>16.4</td>
<td>50.9</td>
<td>35.3</td>
<td>31</td>
</tr>
<tr>
<td>SP – adapt.</td>
<td>20k</td>
<td>48.7</td>
<td>5k</td>
<td>44.3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>PC – adapt.</td>
<td>92k</td>
<td>42.6</td>
<td>11k</td>
<td>32.2</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>BC – adapt.</td>
<td>285k</td>
<td>34.8</td>
<td>31k</td>
<td>19.1</td>
<td>45</td>
<td></td>
</tr>
</tbody>
</table>

Comparing the results of Table 3 and 1, it can be seen that the relative error rate reductions are larger with adapted acoustic models, especially in the SP case. However, the improvements in case of various speech genres are again far from each other.

4.3. Effect of downgraded acoustic conditions

A straightforward idea is to modify not only the language but the acoustic modeling conditions, as well. The aim is to see if the acoustic degradation of planned speech (PC and BC tasks) can reduce the improvement to a similar level as it is measured in the case of spontaneous speech (SP task).

The simplest way of downgrading the acoustic conditions is to add a certain level of white noise to the test speech. Fig. 1 shows the effect of various levels of additive noise on WER’s and relative improvements of downscaled PC and BC tasks.

As it is expected, added noise can decrease the improvements seriously. However, a similar level of error rate reduction as is given in the SP task is reached in the PC and BC case only under unrealistic acoustic conditions.

4.4. Effects of acoustic model adaptation

In this experimental setup we also aim at modifying the acoustic scores, but now in a positive sense: with unsupervised adaptation on the test speech. MLLR (Maximum Likelihood Linear Regression) acoustic model adaptation [13] is performed with the best available systems. Speaker dependent acoustic model transformations were trained only for the SP task, i.e. only one transformation is used per task both in the PC and BC adaptation setups.

Table 3. Results with acoustic model adaptation.

<table>
<thead>
<tr>
<th>Task</th>
<th>Word Vocab</th>
<th>MB WER [%]</th>
<th>-Δrel [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP – adapt.</td>
<td>20k</td>
<td>48.7</td>
<td>9</td>
</tr>
<tr>
<td>PC – adapt.</td>
<td>92k</td>
<td>42.6</td>
<td>24</td>
</tr>
<tr>
<td>BC – adapt.</td>
<td>285k</td>
<td>34.8</td>
<td>45</td>
</tr>
</tbody>
</table>

As Table 2 shows, the larger is the amount of training text data the lower is the recognition error rate. However, the differences of relative improvements between tasks are still preserved with large margins. Table 2 also suggests that the improvements do not depend directly from the OOV rate.
5. Discussion

The results suggest that the high variability of improvements due to morphs-based modeling cannot be explained by either the language model corpus sizes, or by the acoustic conditions. [3] shows that pronunciation modeling does not have a strong influence on morph-based ASR improvement, either.

We suppose that the differences between the examined three speech genres are manifested partially in the different vocabulary growths. Obviously, the number of word forms at a given (sub)corpus size is independent from the previous factors and different for the three LVCSR tasks (see Table 2).

In Fig. 2, besides Hungarian, other language relative improvement results from [4] are shown in the function of number of different word forms at 160k words (sub)corpus sizes. All plotted approaches apply the MB algorithm – though in different ways – and use context dependent phone models. Evidently, test vocabulary growth curves should serve as better basis for such a comparison, but the required information is typically not available, and we assume that test and training data are matched for good recognition accuracies.

The correlation between the number of different word forms and the relative improvements is 0.95 for the Hungarian data set, 0.66 for the whole set, and 0.72 for the whole set but the most outlier TUR1 setup.

6. Conclusions

Three Hungarian LVCSR tasks corresponding to three speech genres were investigated in numerous experimental setups. An important conclusion is that relative error rate reduction – achieved by changing the basis of speech recognition from words to morph – was much less dependent from the type of morphs, from the training text corpus sizes and from the realistic acoustic conditions in our experiments than from the tasks themselves. High correlation was observed between the relative improvements and the number of unique words at a given corpus size for Hungarian tasks. Lower, but still noticeable correlation factors are measured when the investigations were extended to other languages where similar techniques were used. The results can be useful for the speech community if the application of morph- or word-based approach is to be decided for a particular LVCSR task / speech genre.

7. Acknowledgements


8. References