ON MORPH-BASED LVCSR IMPROVEMENTS

Balázs Tarján [1], *Péter Mihajlik* [1,2]

[1] Department of Telecommunication and Media Informatics,
Budapest University of Technology & Economics, Hungary
[2] THINKTech Research Center Nonprofit LLC, Hungary
tarjanb@tmit.bme.hu, mihajlik@tmit.bme.hu

ABSTRACT

Efficient large vocabulary continuous speech recognition of morphologically rich languages is a big challenge due to the rapid vocabulary growth. To improve the results various subword units - called as morphs - are applied as basic language elements. The improvements over the word baseline, however, are changing from negative to error rate halving across languages and tasks. In this paper we make an attempt to explore the source of this variability. Different LVCSR tasks of an agglutinative language are investigated in numerous experiments using full vocabularies. The improvement results are compared to pre-existing other language results, as well. Important correlations are found between the morph-based improvements and between the vocabulary growths and the corpus sizes.

Index Terms — speech recognition, rich morphology, morph, language modeling, LVCSR

1. INTRODUCTION

The most commonly used LVCSR (Large-Vocabulary Continuous Speech Recognition) systems apply words as basic lexical units. Word-based recognition of morphologically rich languages, however, can result in well-known problems [1]: very large vocabularies, high OOV (Out Of Vocabulary) rate, and inaccurate language model parameter estimation due to large number of distinct word forms. These phenomena are handled typically by changing the base units from words to sub-word lexical units called as morphs. In this way vocabulary size can be radically decreased and even OOV words can be recognized. Thus, recognition accuracies can be significantly improved over the word baseline [2-4]. The LVCSR improvement can be outstandingly high in the case of read speech in certain agglutinative languages such as Finnish and Estonian [2]. On the other hand, the reported improvements are much smaller in the case of other agglutinative languages like Turkish, Arabic and Hungarian [4-7]. Besides, improvement for spontaneous speech recognition is very seldom [4] or the results are even worse [8], as compared to the classical word-based approach. Thus, morph-based improvement seems to be not only language but speech genre dependent, as well.

So far, few efforts have been made in order to explain the high variability of improvement due to morph-based speech recognition. A major study compares statistical morphs based LVCSR results across four languages [8]. The difficulty in evaluating these results is that the speech recognition technique was not the same across languages. In [6] the conclusion for Arabic broadcast news recognition is that the improvement of morph-based approach can be eliminated if appropriately large word vocabulary is chosen. [9] also compares morph-based LVCSR to very large vocabulary word-based one but the significant improvements are preserved for Finnish and Estonian. [9] suggests that the relatively worse improvements of others are possibly due to the low order (n<4) of the applied morph n-gram models. All of these work apply empirical cutoffs on the word and morph vocabularies, which make the cross language and across task comparisons difficult. Our previous work was our first attempt to make clear evaluation of morph-based LVCSR across speech genres [10]. It concluded that the vocabulary size at a given training corpus size can be a good indicator for the morph-based improvement. However, some of the Hungarian improvement results were too optimistic due to an unattended cutoff in the word vocabulary and so they became outliers in the across-language comparison.

All in all, in the earlier publications there were always ad-hoc vocabulary cutoffs applied (in the word- or in the morph-based approach or in both cases); therefore the morph and word system comparisons were not entirely fair. In this study, all the results of various LVCSR tasks are measured strictly with full vocabularies. Not only the results of [10] are corrected but important conclusions are sharpened and new ones are found that can be useful for speech recognition of morphologically rich and/or under-resourced languages.

In this paper, the same speech recognition system and the same algorithms are used and optimized separately for three LVCSR tasks: for spontaneous (conversational) speech, for press conference speech, and for classical broadcast news speech. Improvements are measured with well- and less-resourced training text corpora. 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. The conclusions are extended for other languages, as well.

2. TASKS

Since morph-based speech recognition results scatter heavily on a speech genre scale – from read to spontaneous conversational speech – our concept was 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.

2.1. Spontaneous speech – MALACH task (SP)

The Hungarian MALACH task was chosen as the spontaneous end of the scale since no other spontaneous Hungarian database was available for us. The MALACH corpus contains interviews with elderly people and is detailed in [4,11]. The recordings are made typically in normal home environment and their content is carefully transcribed. Only transcriptions are used to train the language models, 160K words in sum. The amount of test data is 4 hours, 19K words (matched data set in [11]).

2.2. 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 80 minutes, 9.4K words.

2.3. Planned speech – Broadcast News task (BN)

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.

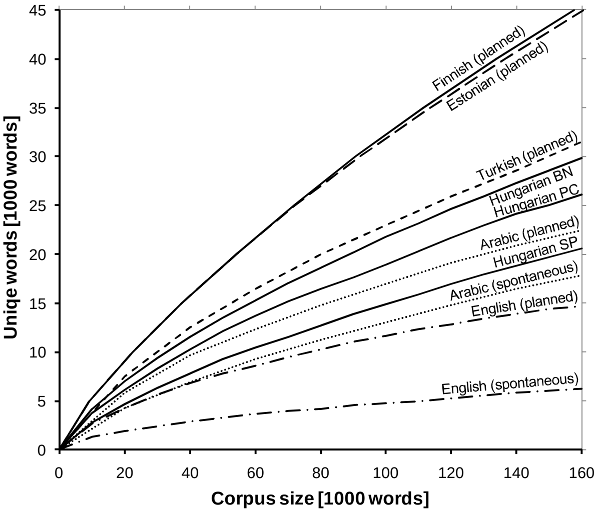

Figure 1: Number of unique words as a function of corpus size. Hungarian curves are calculated on the given databases. Non-Hungarian curves are reproduced from [8] with permission.

The morphological complexities of these tasks can be compared to each other and to other language tasks in Fig. 1.

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.

3. METHODOLOGY

For building morph-based recognizers various vocabulary decomposition algorithms are applied. 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 [12]). 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. The average numbers of morphs composing a word are given in Fig. 2.

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 [13] implemented by the SRILM toolkit [14]. 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 (see Fig. 2). 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 BN corpus, where entropy-based pruning [15] is applied both on the word and morph 3-grams resulting in roughly equal language models in terms of occupied operative memory size.

3.2.2. Pronunciation model

Simple grapheme-to-phoneme rules [16] and exceptions are applied on each lexicon separately in order to obtain word- and morph-to-phoneme mappings. Automatic phonetic transcription of both morphs and words can result in pronunciation errors especially at morph boundaries. However, according to our former experiences this kind of errors occur rarely [4]. Weighted alternative pronunciations are used only for the SP (MALACH) task, though as [4,17] 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 [18]), 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 [18]. Three state left-to-right HMM’s 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 [17]. For both the PC and BN tasks the acoustic models were trained on the MRBA database [19] 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 [20] was also applied.

3.3. Off-line recognition network construction

The WFST (Weighted Finite State Transducer) [21] recognition network is computed on the triphone-level:

$$\text{wred}(\text{fact}(\text{compact}(C \text{ o } S \text{ o } \text{compact}(\text{det}(L \text{ o } G)')))) \quad (1)$$

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 VOXerver – 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 midsized training text corpora for the PC and BN tasks, and about 4 for the largest corpora and for the SP task – measured on the same 3GHz, 1 core CPU.

The SP and the BN test sets contain only the speech of previously unseen speakers. All the PC and BN 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 [8], 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.

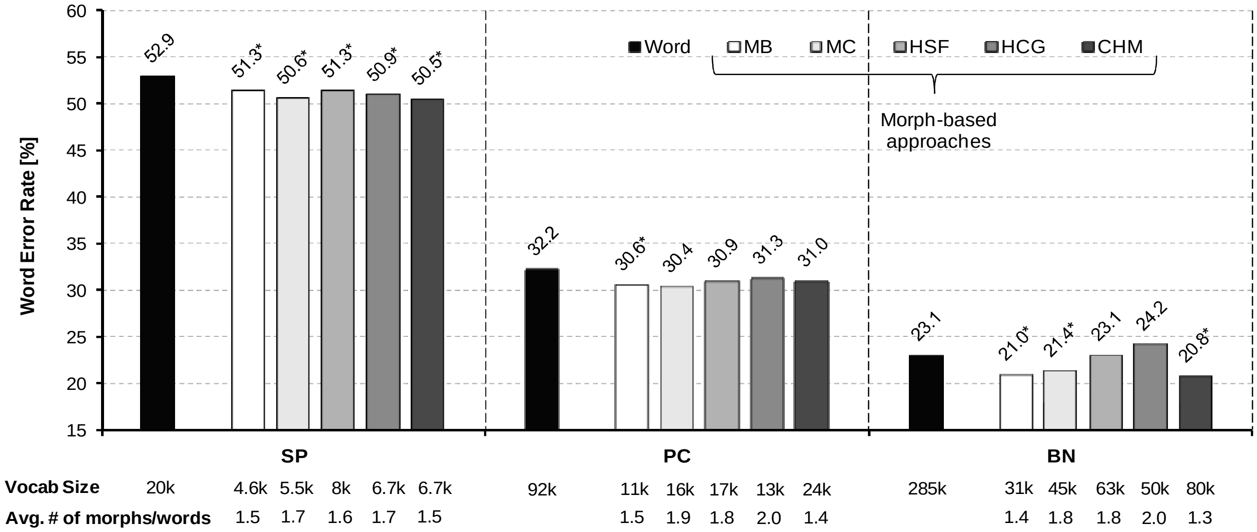

Figure 2: *Full-scale WER results with various morph types*
(* signs indicate significant improvement as compared to word baseline results)

4. RESULTS

First full-scale results are presented on the three Hungarian LVCSR tasks with various morph types. Then, the effect of lower resources – in terms of training text – is investigated. Finally, the influence of acoustic model quality is measured by applying adapted acoustic models.

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, [22]) and MC (Morfessor Categories-MAP, [23]) algorithms.
- ***Grammatical morphs:*** morphs were obtained by applying an affix-stripping method implemented in the Hunmorph system [24,25]. 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 [4,11]

More detailed description of these morph types can be found in [4], [11] and in [17].

As the results in Fig. 2 show, there are significant improvements due to the morph-based LVCSR approach in each task, although the improvements are definitely smaller than the Finnish or Estonian ones [8]. In general, the more planned is the speech the higher is the error rate reduction. Nevertheless, the morph modeling technique does matter, especially in the BN task, where grammatical methods fail to outperform the word baseline. In average, the best results are obtained with the CHM method, but only the MB technique achieves consistently significant word error rate reduction. Moreover, the Morfessor Baseline word-to-morph segmentation method provides the smallest vocabulary sizes, therefore only the MB morph modeling technique is investigated further.

Table 1. Speaker independent speech recognition results with various training text corpora sizes

Task	# of training words	# of word forms	OOV rate [%]	Word WER [%]	MB WER [%]
SP	160k	20k	15.6	52.9	51.3
PC	160k	26k	14.1	43.0	38.4
PC	1.2M	92k	6.3	32.2	30.6
BN	160k	30k	16.4	41.8	35.3
BN	1.2M	105k	7.2	26.4	23.5
BN	5.6M	285k	3.6	23.1	21.0

Table 2. Speech recognition results with acoustic model adaptation and with various training text corpora sizes

Task	# of training words	Word WER [%]	MB WER [%]
SP – adapt.	160k	47.6	43.8
PC – adapt.	160k	40.3	36.0
PC – adapt.	1.2M	30.7	29.1
BN – adapt.	160k	39.8	31.5
BN – adapt.	1.2M	24.3	21.4
BN – adapt.	5.6M	20.6	18.9

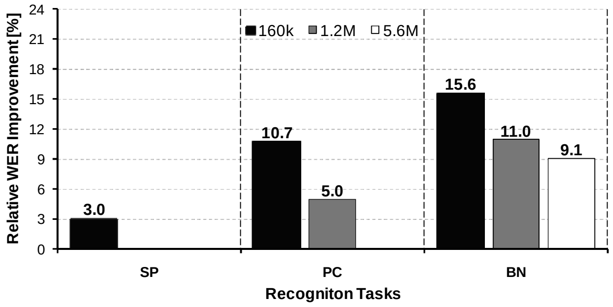

Figure 3: Relative improvement rates of speaker independent WER's measured on different training text corpora sizes

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 BN tasks.

As Table 1 shows – and as it is expected – the larger is the amount of training text data the lower is the recognition error rate. Looking at the improvement results (Fig. 3), a much less expected phenomenon can be observed. Namely, the reduction of training data dramatically increased the improvement rates. This may mean that morph-based speech recognition can be useful tool if the language resources are strongly limited.

By comparing the improvement results of the three tasks measured with equally sized training text it can be seen that there is no direct correspondence between the OOV rates and the improvements (see Table 1 and Fig. 3). The dependence of morph-based error reduction from the speech genre, however, is even more characteristic.

4.3. Effects of acoustic model adaptation

In this experimental setup we aim at modifying the acoustic model quality beside the language model. The previous experiment is repeated using unsupervised MLLR (Maximum Likelihood Linear Regression) acoustic model adaptation instead of applying speaker independent models. 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 BN adaptation setups.

By comparing the results of Table 1 and 2, it can be seen that i) acoustic model adaptation was always effective in the reduction or recognition errors; ii) the relative improvements due to morph-based modeling can be significantly larger with more accurate acoustic models (Fig. 3 and 4).

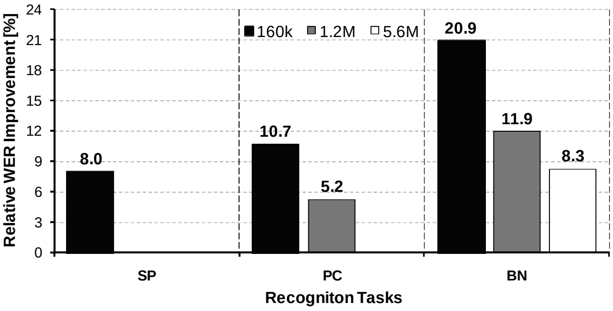

Figure 4: Relative improvement rates of WER's measured with acoustic model adaptation on different training text corpora sizes

5. DISCUSSION

The results suggest that the improvements due to morphs-based modeling correlate greatly with the speech genre. The language dependency assumption is augmented, too: the best improvement for Hungarian is only the half of the best Finnish one [8] achieved with similar techniques.

We suppose that the differences between the examined three speech genres and languages are manifested partially in the different vocabulary growths. Obviously, the number of word forms at a given corpus size is definitely different for the three Hungarian LVCSR tasks as well as for other agglutinative language tasks, see Fig. 1.

In Fig. 5, besides Hungarian, other language relative improvement results from [8] are shown in the function of number of different word forms at 160k words (sub)corpus sizes. All plotted approaches apply the MB algorithm –

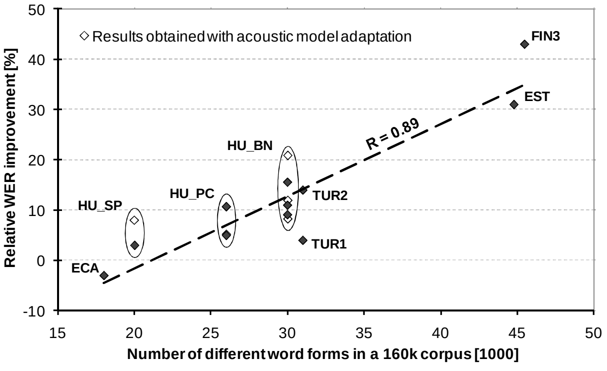

Figure 5: *Illustration of the relation between the relative WER reductions in various language LVCSR setups and the number of unique words at a given amount of training text. (Non-Hungarian results are from [8], ECA stands for spontaneous Arabic, other setups present more or less planned speech.)*

though in different ways – and use context dependent phone models. Although test data is not considered in this comparison 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 0.89 for the whole set, and 0.93 for the whole set but the results obtained with acoustic model adaptation.

6. CONCLUSIONS

Based on the morph-based improvement results of different Hungarian LVCSR tasks and on their comparison to each other and to other agglutinative language results it is possible to draw several conclusions. First, – independently from language and speech genre – the more rapid is the vocabulary growth of the given task the higher improvement can be expected from the application of morph-based speech recognition approach. Second, for vocabulary decomposition, a well-known and publicly available unsupervised statistical method (MB) seems to be a feasible first choice. Furthermore, the results suggest that neither the OOV rate nor the n-gram orders are crucial factor of the improvement, but acoustic model quality may do matter. Finally, an unforeseen conclusion is that morph-based speech recognition can be more beneficial in the case of less resourced tasks. Or, vice versa, using ample data and gigantic word vocabularies may eliminate the error rate reduction due to vocabulary decomposition. To verify and extend this assumption to other languages further researches and resources are needed even though similar phenomenon is observed in [6] evaluating Arabic broadcast news recognition results.

7. ACKNOWLEDGEMENTS

The work was founded partially by national grants: NKFP-2/034/2004, OM-00102-2007, KMOP-1.1.1-07/1-2008-0034, and by TAMOP-4.2.2/08/1/KMR.

8. REFERENCES

[1] K. Kirchoff and R. Sarikaya, “Processing Morphologically-Rich Languages” Tutorial at *INTERSPEECH 2007*, Antwerp, Belgium

[2] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimaki, J. Pylkkonen, T. Alumae and M. Saraclar, “Unlimited vocabulary speech recognition for agglutinative languages,” in *HLT-NAACL*, New York, USA, June 5-7, 2006.

[3] O.-W. Kwon and J. Park, “Korean large vocabulary continuous speech recognition with morpheme-based recognition units,” *Speech Communication*, vol. 39, Issue 3-4, pp. 287-300, Feb. 2003.

[4] P. Mihajlik, Z. Tüske, B. Tarján, B. Németh and T. Fegyó, “Improved recognition of spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task,” *IEEE Transactions on Audio, Speech, and Language Processing* – in press

[5] E. Arisoy, D. Can, S. Parlak, H. Sak and M. Saraclar, “Turkish Broadcast News Transcription and Retrieval,” *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):874-883, July 2009.

[6] G. Choueiter, D. Povey, S.F. Chen, and G. Zweig, ”Morpheme-based language modeling for Arabic LVCSR,” in *Proc. ICASSP’06*, Tolouse, France, 2006.

[7] M. Afify, R. Sarikaya, H.-K. J. Kuo, L. Besacier and Y. Gao, “On the Use of Morphological Analysis for Dialectal Arabic Speech Recognition,” in *INTERSPEECH-2006*, paper 1444.

[8] M. Creutz et. al., “Morph-Based Speech Recognition and Modeling Out-of-Vocabulary Words Across Languages,” *ACM Transactions on Speech and Language Processing*, vol. 5, Issue 1, Article no. 3, December 2007.

[9] T. Hirsimäki, J. Pylkkönen and M. Kurimo, „Importance of High-Order N-gram Models in Morph-Based Speech Recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, 17(4): 724-732, May 2009.

[10] P. Mihajlik, B. Tarján, Z. Tüske and T. Fegyó, “Investigation of Morph-based Speech Recognition Improvements across Speech Genres,” in *Proc. Interspeech*, Brighton, United Kingdom, 2009, pp. 2687-2690.

[11] P. Mihajlik, T. Fegyó, B. Németh, Z. Tüske and V. Trón, “Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages,” in *TSD 2007*, Pilsen, Czech Republic, September 2007.

[12] T. Hirsimaki and M. Kurimo, “Decoder issues in unlimited Finnish speech recognition” In Proceedings of the Nordic Signal Processing Symposium *NORSIG 2004*, Espoo, Finland, 2004.

[13] S.F. Chen and J.T. Goodman, ”An Empirical Study of Smoothing Techniques for Language Modeling” Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[14] A. Stolcke, “SRILM – an extensible language modeling toolkit,” in *Proc. Intl. Conf. on Spoken Language Processing*, pp. 901–904, Denver, 2002.

[15] A. Stolcke, “Entropy-based pruning of backoff language models,” in *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, 270–274

[16] M. Szarvas, T. Fegyó, P. Mihajlik and P. Tatai, “Automatic Recognition of Hungarian: Theory and Practice,” *International Journal of Speech Technology*, 3:277-287, December 2000.

[17] P. Mihajlik, T. Fegyó, Z. Tüske and P. Ircing, “A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – Like Hungarian,” in *INTERSPEECH-2007*, pp. 1497-1500.

[18] S. Young, D. Ollason, V. Valtchev and P. Woodland, *The HTK book.* (for HTK version 3.2.), March 2002.

[19] MRBA – Hungarian Language Speech Database, http://alpha.tmit.bme.hu/speech/hdbMRBA.php

[20] L. Mauuary. “Blind Equalization in the Cepstral Domain for robust Telephone based Speech Recognition,” in *Proc. of EUSPICO’98*, Vol.1, pp. 359-363, 1998.

[21] M. Mohri, F. Pereira and M. Riley, “Weighted Finite-State Transducers in Speech Recognition,” *Computer Speech and Language*, 16(1):69-88, 2002.

[22] M. Creutz and K. Lagus, “Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.,” in *Comp. and Inf. Sci.*, report A81, March 2005.

[23] M. Creutz and K. Lagus, “Inducing the Morphological Lexicon of a Natural Language from Unannotated Text,” in *Proc. of AKRR'05*, *Espoo*, Finland, 15-17 June, 2005.

[24] V. Trón, L. Németh, P. Halácsy, A. Kornai, Gy. Gyepesi and D. Varga, “Hunmorph: open source word analysis,” in *Proc. ACL 2005 Software Workshop*, pp. 77-85.

[25] V. Trón, P. Halácsy, P. Rebrus, A. Rung, E. Simon and P. Vajda, “morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis,” (in Hungarian) in *MSZNY Conf.*, Szeged, 2005.