Unlimited vocabulary speech recognition for agglutinative languages

Mikko Kurimo[1], **Antti Puurula**[1], **Ebru Arisoy**[2], **Vesa Siivola**[1],
Teemu Hirsimäki[1], **Janne Pylkkönen**[1], **Tanel Alumäe**[3], **Murat Saraclar**[2]
[1] Adaptive Informatics Research Centre, Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Finland
{Mikko.Kurimo,Antti.Puurula,Vesa.Siivola}@tkk.fi
[2] Bogazici University, Electrical and Electronics Eng. Dept.
34342 Bebek, Istanbul, Turkey
{arisoyeb,murat.saraclar}@boun.edu.tr
[3] Laboratory of Phonetics and Speech Technology,
Institute of Cybernetics, Tallinn Technical University, Estonia
tanel.alumae@phon.ioc.ee

Abstract

It is practically impossible to build a word-based lexicon for speech recognition in agglutinative languages that would cover all the relevant words. The problem is that words are generally built by concatenating several prefixes and suffixes to the word roots. Together with compounding and inflections this leads to millions of different, but still frequent word forms. Due to inflections, ambiguity and other phenomena, it is also not trivial to automatically split the words into meaningful parts. Rule-based morphological analyzers can perform this splitting, but due to the handcrafted rules, they also suffer from an out-of-vocabulary problem. In this paper we apply a recently proposed fully automatic and rather language and vocabulary independent way to build subword lexica for three different agglutinative languages. We demonstrate the language portability as well by building a successful large vocabulary speech recognizer for each language and show superior recognition performance compared to the corresponding word-based reference systems.

1 Introduction

Speech recognition for dictation or prepared radio and television broadcasts has had huge advances during the last decades. For example, broadcast news (BN) in English can now be recognized with about ten percent word error rate (WER) (NIST, 2000) which results in mostly quite understandable text. Some rare and new words may be missing but the result has proven to be sufficient for many important applications, such as browsing and retrieval of recorded speech and information retrieval from the speech (Garofolo et al., 2000). However, besides the development of powerful computers and new algorithms, a crucial factor in this development is the vast amount of transcribed speech and suitable text data that has been collected for training the models. The problem faced in porting the BN recognition systems to conversational speech or to other languages is that almost as much new speech and text data have to be collected again for the new task.

The reason for the need for a vast amount of training texts is that the state-of-the-art statistical language models contain a huge amount of parameters to be estimated in order to provide a proper probability for any possible word sequence. The main reason for the huge model size is that for an acceptable coverage in an English BN task, the vocabulary must be very large, at least 50,000 words, or more. For languages with a higher degree of word inflections than English, even larger vocabularies are required. This paper focuses on the agglutinative languages in which words are frequently formed by concatenating one or more stems, prefixes, and suffixes. For these languages in which the words are often highly inflected as well as formed from several morphemes, even a vocabulary of 100,000 most common words would not give sufficient coverage (Kneissler and

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 487–494,
New York, June 2006.

Klakow, 2001; Hirsimäki et al., 2005). Thus, the solution to the language modeling clearly has to involve splitting of words into smaller modeling units that could then be adequately modeled.

This paper focuses on solving the vocabulary problem for several languages in which the speech and text database resources are much smaller than for the world's main languages. A common feature for the agglutinative languages, such as Finnish, Estonian, Hungarian and Turkish is that the large vocabulary continuous speech recognition (LVCSR) attempts so far have not resulted comparable performance to the English systems. The reason for this is not only the language modeling difficulties, but, of course, the lack of suitable speech and text training data resources. In (Geutner et al., 1998; Siivola et al., 2001) the systems aim at reducing the active vocabulary and language models to a feasible size by clustering and focusing. In (Szarvas and Furui, 2003; Alumäe, 2005; Hacioglu et al., 2003) the words are split into morphemes by language-dependent hand-crafted morphological rules. In (Kneissler and Klakow, 2001; Arisoy and Arslan, 2005) different combinations of words, grammatical morphemes and endings are utilized to decrease the OOV rate and optimize the speech recognition accuracy. However, constant large improvements over the conventional word-based language models in LVCSR have been rare.

The approach presented in this paper relies on a data-driven algorithm called Morfessor (Creutz and Lagus, 2002; Creutz and Lagus, 2005) which is a language independent unsupervised machine learning method to find morpheme-like units (called statistical morphs) from a large text corpus. This method has several advantages over the rule-based grammatical morphemes, e.g. that no hand-crafted rules are needed and all words can be processed, even the foreign ones. Even if good grammatical morphemes are available, the language modeling results by the statistical morphs seem to be at least as good, if not better (Hirsimäki et al., 2005). In this paper we evaluate the statistical morphs for three agglutinative languages and describe three different speech recognition systems that successfully utilize the n-gram language models trained for these units in the corresponding LVCSR tasks.

2 Building the lexicon and language models

2.1 Unsupervised discovery of morph units

Naturally, there are many ways to split the words into smaller units to reduce a lexicon to a tractable size. However, for a subword lexicon suitable for language modeling applications such as speech recognition, several properties are desirable:

1. The size of the lexicon should be small enough that the n-gram modeling becomes more feasible than the conventional word based modeling.
2. The coverage of the target language by words that can be built by concatenating the units should be high enough to avoid the out-of-vocabulary problem.
3. The units should be somehow meaningful, so that the previously observed units can help in predicting the next one.
4. In speech recognition one should be able to determine the pronunciation for each unit.

A common approach to find the subword units is to program the language-dependent grammatical rules into a morphological analyzer and utilize that to then split the text corpus into morphemes as in e.g. (Hirsimäki et al., 2005; Alumäe, 2005; Hacioglu et al., 2003). There are some problems related to ambiguous splits and pronunciations of very short inflection-type units, but also the coverage in, e.g., news texts may be poor because of many names and foreign words.

In this paper we have adopted a similar approach as (Hirsimäki et al., 2005). We use unsupervised learning to find the best units according to some cost function. In the Morfessor algorithm the minimized cost is the coding length of the lexicon and the words in the corpus represented by the units of the lexicon. This minimum description length based cost function is especially appealing, because it tends to give units that are both as frequent and as long as possible to suit well for both training the language models and also decoding of the speech. Full coverage of the language is also guaranteed by splitting the rare words into very short units, even to single phonemes if necessary. For language models utilized in speech recognition, the lexicon of the statistical morphs can be further reduced by omitting the rare words from the input of the Morfessor algorithm. This operation does not reduce the coverage of the lexicon, because it just splits the rare words then into smaller units, but the smaller lexicon may offer a remarkable speed up of the recognition.

The pronunciation of, especially, the short units may be ambiguous and may cause severe problems in languages like English, in which the pronunciations can not be adequately determined from the orthography. In most agglutinative languages, such as Finnish, Estonian and Turkish, rather simple letter-to-phoneme rules are, however, sufficient for most cases.

2.2 Building the lexicon for open vocabulary

The whole training text corpus is first passed through a word splitting transformation as in Figure 1. Based on the learned subword unit lexicon, the best split for each word is determined by performing a Viterbi search with the unigram probabilities of the units. At this point the word break symbols are added between each word in order to incorporate that information in the statistical language models, as well. Then the n-gram models are trained similarly as if the language units were words including word and sentence break symbols as additional units.

2.3 Building the n-gram model over morphs

Even though the required morph lexicon is much smaller than the lexicon for the corresponding word n-gram estimation, the data sparsity problem is still important. Interpolated Kneser-Ney smoothing is utilized to tune the language model probabilities in the same way as found best for the word n-grams. The n-grams that are not very useful for modeling the language can be discarded from the model in order to keep the model size down. For Turkish, we used the entropy based pruning (Stolcke, 1998), where the n-grams, that change the model entropy less than a given treshold, are discarded from the model. For Finnish and Estonian, we used n-gram growing (Siivola and Pellom, 2005). The n-grams that increase the training set likelihood enough with respect to the corresponding increase in the model size are accepted into the model (as in the minimum description length principle). After the growing process the model is further pruned with entropy based pruning. The method allows us to train models with higher order n-grams, since the memory consumption is lower and also gives somewhat better models. Both methods can also be viewed as choosing the correct model complexity for the training data to avoid over-learning.

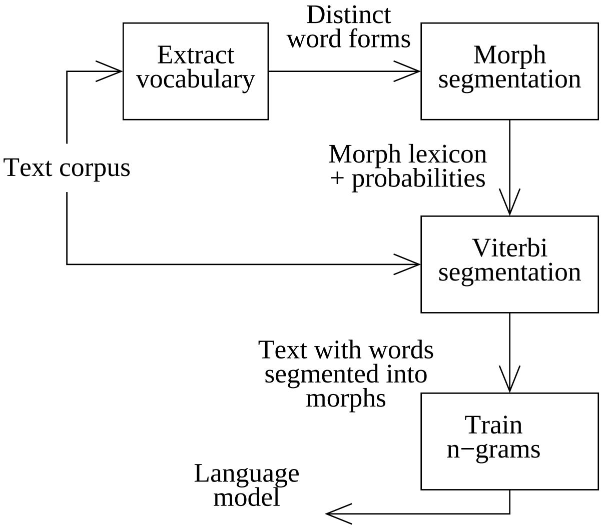

Figure 1: The steps in the process of estimating a language model based on statistical morphs from a text corpus (Hirsimäki et al., 2005).

3 Statistical properties of Finnish, Estonian and Turkish

Before presenting the speech recognition results, some statistical properties are presented for the three agglutinative languages studied. If we consider choosing a vocabulary of the 50k-70k most common words, as usual in English broadcast news LVCSR systems, the out-of-vocabulary (OOV) rate in English is typically smaller than 1%. Using the language model training data the following OOV rates can be found for a vocabulary including only the most common words: 15% OOV for 69k in Finnish (Hirsimäki et al., 2005), 10% for 60k in Estonian and 9% for 50k in Turkish. As shown in (Hacioglu et al., 2003) this does not only mean the same amount of extra speech recognition errors, but even more, because the recognizer tends to lose track when unknown words get mapped to those that are in the vocabulary. Even doubling the vocabulary is not a suf-

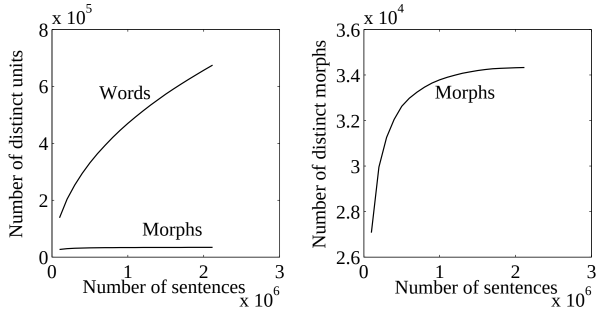

Figure 2: Vocabulary growth of words and morphs for Turkish language

ficient solution, because a vocabulary twice as large (120k) would only reduce the OOV rate to 6% in Estonian and 5% in Turkish. In Finnish even a 400k vocabulary of the most common words still gives 5% OOV in the language model training material.

Figure 2 illustrates the vocabulary explosion encountered when using words and how using morphs avoids this problem for Turkish. The figure on the left shows the vocabulary growth for both words and morphs. The figure on the right shows the graph for morphs in more detail. As seen in the figure, the number of new words encountered continues to increase as the corpus size gets larger whereas the number of new morphs encountered levels off.

4 Speech recognition experiments

4.1 About selection of the recognition tasks

In this work the morph-based language models have been applied in speech recognition for three different agglutinative languages, Finnish, Estonian and Turkish. The recognition tasks are speaker dependent and independent fluent dictation of sentences taken from newspapers and books, which typically require very large vocabulary language models.

4.2 Finnish

Finnish is a highly inflected language, in which words are formed mainly by agglutination and compounding. Finnish is also the language for which the algorithm for the unsupervised morpheme discovery (Creutz and Lagus, 2002) was originally developed. The units of the morph lexicon for the experiments in this paper were learned from a joint corpus containing newspapers, books and newswire stories of totally about 150 million words (CSC, 2001). We obtained a lexicon of 25k morphs by feeding the learning algorithm with the word list containing the 160k most common words. For language model training we used the same text corpus and the recently developed growing n-gram training algorithm (Siivola and Pellom, 2005). The amount of resulted n-grams are listed in Table 4. The average length of a morph is such that a word corresponds to 2.52 morphs including a word break symbol.

The speech recognition task consisted of a book read aloud by one female speaker as in (Hirsimäki et al., 2005). Speaker dependent cross-word triphone models were trained using the first 12 hours of data and evaluated by the last 27 minutes. The models included tied state hidden Markov models (HMMs) of totally 1500 different states, 8 Gaussian mixtures (GMMs) per state, short-time mel-cepstral features (MFCCs), maximum likelihood linear transformation (MLLT) and explicit phone duration models (Pylkkönen and Kurimo, 2004). The real-time factor of recognition speed was less than 10 xRT with a 2.2 GHz CPU. However, with the efficient LVCSR decoder utilized (Pylkkönen, 2005) it seems that by making an even smaller morph lexicon, such as 10k, the decoding speed could be optimized to only a few times real-time without an excessive trade-off with recognition performance.

4.3 Estonian

Estonian is closely related to Finnish and a similar language modeling approach was directly applied to the Estonian recognition task. The text corpus used to learn the morph units and train the statistical language model consisted of newspapers and books, altogether about 55 million words (Segakorpus, 2005). At first, 45k morph units were obtained as the best subword unit set from the list of the 470k most common words in the corpora. For speeding up the recognition, the morph lexicon was afterwards reduced to 37k by splitting the rarest morphs (occurring in only one or two words) further into smaller ones. Corresponding growing n-gram language models as in Finnish were trained from the Estonian corpora resulting the n-grams in Table 4.

The speech recognition task in Estonian consisted of long sentences read by 50 randomly picked held-out test speakers, 7 sentences each (a part of (Meister et al., 2002)). Unlike the Finnish and Turkish microphone data, this data was recorded from telephone, i.e. 8 kHz sampling rate and narrow band data instead of 16 kHz and normal (full) bandwidth. The phoneme models were trained for speaker independent recognition using windowed cepstral mean subtraction and significantly more data (over 200 hours and 1300 speakers) than for the Finnish task. The speaker independence, together with the telephone quality and occasional background noises, made this task still a considerably more difficult one. Otherwise the acoustic models were similar cross-word triphone GMM-HMMs with MFCC features, MLLT transformation and the explicit phone duration modeling, except larger: 5100 different states and 16 GMMs per state. Thus, the recognition speed is also slower than in Finnish, about 20 xRT (2.2GHz CPU).

4.4 Turkish

Turkish is another a highly-inflected and agglutinative language with relatively free word order. The same Morfessor tool (Creutz and Lagus, 2005) as in Finnish and Estonian was applied to Turkish texts as well. Using the 360k most common words from the training corpus, 34k morph units were obtained. The training corpus consists of approximately 27M words taken from literature, law, politics, social sciences, popular science, information technology, medicine, newspapers, magazines and sports news. N-gram language models for different orders with interpolated Kneser-Ney smoothing as well as entropy based pruning were built for this morph lexicon using the SRILM toolkit (Stolcke, 2002). The number of n-grams for the highest order we tried (6-grams without entropy-based pruning) are reported in Table 4. In average, there are 2.37 morphs per word including the word break symbol.

The recognition task in Turkish consisted of approximately one hour of newspaper sentences read by one female speaker. We used decision-tree state clustered cross-word triphone models with approximately 5000 HMM states. Instead of using letter to phoneme rules, the acoustic models were based directly on letters. Each state of the speaker independent HMMs had a GMM with 6 mixture components. The HTK frontend (Young et al., 2002) was used to get the MFCC based acoustic features. The explicit phone duration models were not applied. The training data contained 17 hours of speech from over 250 speakers. Instead of the LVCSR decoder used in Finnish and Estonian (Pylkkönen, 2005), the Turkish evaluation was performed using another decoder (AT&T, 2003), Using a 3.6GHz CPU, the real-time factor was around one.

5 Results

The recognition results for the three different tasks: Finnish, Estonian and Turkish, are provided in Tables 1 – 3. In each task the word error rate (WER) and letter error rate (LER) statistics for the morph-based system is compared to a corresponding word-based system. The resulting morpheme strings are glued to words according to the word break symbols included in the language model (see Section 2.2) and the WER is computed as the sum of substituted, inserted and deleted words divided by the correct number of words. LER is included here as well, because although WER is a more common measure, it is not comparable between languages. For example, in agglutinative languages the words are long and contain a variable amount of morphemes. Thus, any incorrect prefix or suffix would make the whole word incorrect. The n-gram language model statistics are given in Table 4.

Finnish	lexicon	WER	LER
Words	400k	8.5	1.20
Morphs	25k	7.0	0.95

Table 1: The LVCSR performance for the speaker-dependent Finnish task consisting of book-reading (see Section 4.2). For a reference (word-based) language model a 400k lexicon was chosen.

Estonian	lexicon	WER	LER
Words	60k	56.3	22.4
Morphs	37k	47.6	18.9

Table 2: The LVCSR performance for the speaker-independent Estonian task consisting of read sentences recorded via telephone (see Section 4.3). For a reference (word-based) language model a 60k lexicon was used here.

Turkish	lexicon	WER	LER
Words			
3-gram	50k	38.8	15.2
Morphs			
3-gram	34k	39.2	14.8
4-gram	34k	35.0	13.1
5-gram	34k	33.9	12.4
Morphs, rescored by morph 6-gram			
3-gram	34k	33.8	12.4
4-gram	34k	33.2	12.3
5-gram	34k	33.3	12.2

Table 3: The LVCSR performance for the speaker-independent Turkish task consisting of read newspaper sentences (see Section 4.4). For the reference 50k (word-based) language model the accuracy given by 4 and 5-grams did not improve from that of 3-grams.

#	morph-based models		
ngrams	Finnish	Estonian	Turkish
1grams	24,833	37,061	34,332
2grams	2,188,476	1,050,127	655,621
3grams	17,064,072	7,133,902	1,936,263
4grams	25,200,308	8,201,543	3,824,362
5grams	7,167,021	3,298,429	4,857,125
6grams	624,832	691,899	5,523,922
7grams	23,851	55,363	-
8grams	0	1045	-
Sum	52,293,393	20,469,369	16,831,625

Table 4: The amount of different n-grams in each language model based on statistical morphs. Note that the Turkish language model was not prepared by the growing n-gram algorithm as the others and the model was limited to 6-grams.

In the Turkish recognizer the memory constraints during network optimization (Allauzen et al., 2004) allowed the use of language models only up to 5-grams. The language model pruning thresholds were optimized over a range of values and the best results are shown in Table 3. We also tried the same experiments with two-pass recognition. In the first pass, instead of the best path, lattice output was generated with the same language models with pruning. Then these lattices were rescored using the non-pruned 6-gram language models (see Table 4) and the best path was taken as the recognition output. For the word-based reference model, the two-pass recognition gave no improvements. It is likely that the language model training corpus was too small to train proper 6-gram word models. However, for the morph-based model, we obtained a slight improvement (0.7 % absolute) by two-pass recognition.

6 Discussion

The key result of this paper is that we can successfully apply the unsupervised statistical morphs in large vocabulary language models in all the three experimented agglutinative languages. Furthermore, the results show that in all the different LVCSR tasks, the morph-based language models perform very well and constantly dominate the reference language model based on words. The way that the lexicon is built from the word fragments allows the construction of statistical language models, in practice, for almost an unlimited vocabulary by a lexicon that still has a convenient size.

The recognition was here restricted to agglutinative languages and tasks in which the language used is both rather general and matches fairly well with the available training texts. Significant performance variation in different languages can be observed here, because of the different tasks and the fact that comparable recognition conditions and training resources have not been possible to arrange. However, we believe that the tasks are still both difficult and realistic enough to illustrate the difference of performance when using language models based on a lexicon of morphs vs. words in each task. There are no directly comparable previous LVCSR results on the same tasks and data, but the closest ones which can be found are slightly over 20% WER for the Finnish task (Hirsimäki et al., 2005), slightly over 40 % WER for the Estonian task (Alumäe, 2005) and slightly over 30 % WER for the Turkish task (Erdogan et al., 2005).

Naturally, it is also possible to prepare a huge lexicon and still succeed in recognition fairly well (Saraclar et al., 2002; McTait and Adda-Decker, 2003; Hirsimäki et al., 2005), but this is not a very convenient approach because of the resulting huge language models or the heavy pruning required to keep them still tractable. The word-based language models that were constructed in this paper as reference models were trained as much as possible in the same way as the corresponding morph language models. For Finnish and Estonian the growing n-grams (Siivola and Pellom, 2005) were used including the option of constructing the OOV words from phonemes as in (Hirsimäki et al., 2005). For Turkish a conventional n-gram was built by SRILM similarly as for the morphs. The recognition approach taken for Turkish involves a static decoding network construction and optimization resulting in near real time decoding. However, the memory requirements of network optimization becomes prohibitive for large lexicon and language models as presented in this paper.

In this paper the recognition speed was not a major concern, but from the application point of view that is a very important factor to be taken into a account in the comparison. It seems that the major factors that make the recognition slower are short lexical units, large lexicon and language models and the amount of Gaussian mixtures in the acoustic model.

7 Conclusions

This work presents statistical language models trained on different agglutinative languages utilizing a lexicon based on the recently proposed unsupervised statistical morphs. To our knowledge this is the first work in which similarly developed subword unit lexica are developed and successfully evaluated in three different LVCSR systems in different languages. In each case the morph-based approach constantly shows a significant improvement over a conventional word-based LVCSR language models. Future work will be the further development of also the grammatical morph-based language models and comparison of that to the current approach, as well as extending this evaluation work to new languages.

8 Acknowledgments

We thank the Finnish Federation of the Visually Impaired for providing the Finnish speech data and the Finnish news agency (STT) and the Finnish IT center for science (CSC) for the text data. Our work was supported by the Academy of Finland in the projects *New information processing principles*, *Adaptive Informatics* and *New adaptive and learning methods in speech recognition*. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. The authors would like to thank Sabanci and ODTU universities for the Turkish acoustic and text data and AT&T Labs – Research for the software. This research is partially supported by SIMILAR Network of Excellence and TUBITAK BDP (Unified Doctorate Program of the Scientific and Technological Research Council of Turkey).

References

Cyril Allauzen, Mehryar Mohri, Michael Riley, and Brian Roark. 2004. A generalized construction of integrated speech recognition transducers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada.

Tanel Alumäe. 2005. Phonological and morphological modeling in large vocabulary continuous Estonian speech recognition system. In *Proceedings of Second Baltic Conference on Human Language Technologies*, pages 89–94.

Mehryar Mohri and Michael D. Riley. DCD Library – Speech Recognition Decoder Library. AT&T Labs – Research. `http://www.research.att.com/sw/tools/dcd/`.

Ebru Arisoy and Levent Arslan. 2005. Turkish dictation system for broadcast news applications. In *13th European Signal Processing Conference - EUSIPCO 2005*, Antalya, Turkey, September.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology. URL: `http://www.cis.hut.fi/projects/morpho/`.

J. Garofolo, G. Auzanne, and E. Voorhees. 2000. The TREC spoken document retrieval track: A success story. In *Proceedings of Content Based Multimedia Information Access Conference*, April 12-14.

P. Geutner, M. Finke, and P. Scheytt. 1998. Adaptive vocabularies for transcribing multilingual broadcast news. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, USA, May.

H. Erdogan, O. Buyuk, K. Oflazer. 2005. Incorporating language constraints in sub-word based speech recognition. IEEE Automatic Speech Recognition and Understanding Workshop, Cancun, Mexico.

Kadri Hacioglu, Brian Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo, and Mathias Creutz. 2003. On lexicon creation for Turkish LVCSR. In *Proceedings of 8th European Conference on Speech Communication and Technology*, pages 1165–1168.

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. 2005. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*. (accepted for publication).

Jan Kneissler and Dietrich Klakow. 2001. Speech recognition for huge vocabularies by using optimized subword units. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 69–72, Aalborg, Denmark.

CSC Tieteellinen laskenta Oy. 2001. Finnish Language Text Bank: Corpora Books, Newspapers, Magazines and Other. `http://www.csc.fi/kielipankki/`.

Kevin McTait and Martine Adda-Decker. 2003. The 300k LIMSI German Broadcast News Transcription System. In *Proceedings of 8th European Conference on Speech Communication and Technology*.

Einar Meister, Jürgen Lasn, and Lya Meister. 2002. Estonian SpeechDat: a project in progress. In *Proceedings of the Fonetiikan Päivät – Phonetics Symposium 2002 in Finland*, pages 21–26.

NIST. 2000. *Proceedings of DARPA workshop on Automatic Transcription of Broadcast News*. NIST, Washington DC, May.

Janne Pylkkönen. 2005. New pruning criteria for efficient decoding. In *Proceedings of 9th European Conference on Speech Communication and Technology*.

Janne Pylkkönen and Mikko Kurimo. 2004. Duration modeling techniques for continuous speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*.

Murat Saraclar, Michael Riley, Enrico Bocchieri, and Vincent Goffin. 2002. Towards automatic closed captioning: Low latency real time broadcast news transcription. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, USA.

Segakorpus – Mixed Corpus of Estonian. Tartu University. `http://test.cl.ut.ee/korpused/segakorpus/`.

Vesa Siivola and Bryan Pellom. 2005. Growing an n-gram language model. In *Proceedings of 9th European Conference on Speech Communication and Technology*.

Vesa Siivola, Mikko Kurimo, and Krista Lagus. 2001. Large vocabulary statistical language modeling for continuous speech recognition. In *Proceedings of 7th European Conference on Speech Communication and Technology*, pages 737–747, Aalborg, Copenhagen.

Andreas Stolcke. 1998. Entropy-based pruning of backoff language models. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904.

Mate Szarvas and Sadaoki Furui. 2003. Evaluation of the stochastic morphosyntactic language model on a one million word Hungarian task. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, pages 2297–2300.

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