Large Vocabulary Continuous Speech Recognition for Estonian Using Morpheme Classes

Tanel Alumäe

Laboratory of Phonetics and Speech Technology
Institute of Cybernetics
Tallinn Technical University, Estonia
tanel.alumae@phon.ioc.ee

Abstract

This paper describes development of a large vocabulary continuous speaker independent speech recognition system for Estonian. Estonian is an agglutinative language and the number of different word forms is very large, in addition, the word order is relatively unconstrained. To achieve a good language coverage, we use pseudo-morphemes as basic units in a statistical trigram language model. To improve language model robustness, we automatically find morpheme classes and interpolate the morpheme model with the class-based model. The language model is trained on a newspaper corpus of 15 million word forms. Clustered triphones with multiple Gaussian mixture components are used for acoustic modeling. The system with interpolated morpheme language model is found to perform significantly better than the baseline word form trigram system in all areas. The word error rate of the best system is 27.3% which is a 10.0% absolute improvement over the baseline system.

1. Introduction

The objective of our work is to build a large vocabulary continuous speech recognition system for Estonian. For languages like English, German and French, many successful large vocabulary speech recognition systems have been developed and commercial systems are widely available. In spite of active research in the area of phonetics and computational linguistics, there has been only minor attempts for developing Estonian large vocabulary speech recognition. There are two main reasons for this. First, the number of native Estonians is only about one million, thus there is little interest from commercial organizations for active research and development in the area. Second, Estonian is an agglutinative language, thus it's words are heavily inflected depending on their syntactic role. This makes the number of distinctive words in the language very large. Therefore, a high out-of-vocabulary (OOV) rate is expected when words are used as recognition units in composing a statistical language model [1]. Also, the word order in Estonian is much more free than in non-agglutinative languages like English, which also complicates building of a good language model.

Recently, promising speech recognition results for many agglutinative languages have been reported. In order to increase vocabulary coverage, subword units are used as basic units in language modeling. Subword units may be found using morphological analysis [2, 3], or discovered automatically based on some criteria [4].

In this paper, we describe development of large vocabulary speech recognition that uses words, morphemes and their classes for language modeling. Inflected words are split (with some constraints) into stems and endings using a morphoanalytical tool. The resulting units are automatically clustered into classes in order to increase language model robustness. Speech recognition experiments with interpolated class and morpheme based language model show a significant drop in word error rate over the baseline word form and over the morpheme based system.

2. Resources

The only systematically composed speech database for Estonian that was available when the research started is the Estonian Phonetic Database which is part of the BABEL multi-language database [5]. The corpus is partitioned into 3 sets: many talker set (30 male and 30 female speakers), few talker set (4 male and 4 female speakers) and a very few talker set (a male and a female speaker). Speech recordings have been performed in an anechoic room and digitized at 20 kHz and 16 bit. For each speaker, there are one or more recorded text passages, a set of isolated utterances, and a set of isolated read numbers. The texts that were read were selected so that all main phonologically relevant oppositions would be revealed in the corpus. All recordings come with sentence level transcriptions in both orthographic as well as SAMPA phonemic formats. In addition to the mentioned recordings, there are also isolated CVC construction recordings in the database, which were however not used for training the recognizer.

The text passage and sentence recordings in the many talker set were used for training the acoustic models. This resulted in 1230 audio files, approximately 3:45 of duration. The isolated sentence recordings in the few talker set, approximately 16 minutes of speech, were used for evaluating the performance of the speech recognizer.

For language model training, a part of the Tartu University corpus of Estonian literary language was used [6]. The used part contains approximately 15 million words. Most of the texts come from two national newspapers, “Postimees” and “Eesti Ekspress”, and only about 5% from fiction literature written in 1990s.

For language model evaluation, the texts of the BABEL speech database were used. The texts are relatively neutral in style, resembling more fiction than newspaper articles.

3. Properties of the Estonian language

Estonian belongs to the group of fenno-ugric languages and is thus heavily inflected. One or many suffixes can be appended to verb and noun stems, depending on their syntactic and semantic role in the sentence. Compound words are written together. For example, the word “Eesti-keelse-te-ga-gi” can be translated as “Even with those in the Estonian language”.

The word order in Estonian is much more free than in non-agglutinative languages like English. In more simple sentences, word order can be interchanged without changing the meaning of the sentence, although in general, a typical word order exists and the alternations are used to express some kind of meta-information (e.g. stress).

4. LVCSR system

4.1. Acoustic modeling

The acoustic models were trained using the Hidden Markov Model Toolkit (HTK) [7], version 3.2.

For feature extraction, we use a 25 ms Hamming window with a shift of 10 ms. Each feature vector consists of 13 MFCC coefficients together with the corresponding delta and acceleration coefficients. Cepstral mean normalization calculated for each sentence separately is applied to the coefficients.

The system’s basic units of recognition are phonemes which are modeled by hidden Markov models (HMMs). 22 phonemes, a silence and a a possible short pause between words are modeled. Phonemes of long and over-long duration are modeled as sequences of two short duration models. Phoneme models have three emitting states and a left-to-right topology. Monophone models are bootstrapped from flat start, cloned into triphones and state-clustered using a phonetic decision tree. Finally, eight Gaussian mixture components for each physical state are trained. Inter-morpheme and inter-word modeling is used.

The system dictionary is created automatically from pseudomorpheme orthography. Estonian orthography is not entirely phonetic but simple experiments to model deviations from the orthographic form did not yield higher recognition accuracy, on the contrary, the scores were lower. More complex experiments were left for future work.

4.2. Language modeling

The agglutinativeness of the Estonian language implies that the total number of word forms is huge. Thus, a word n-gram model of 65 000 most frequent words, that is the most common statistical language model in speech recognition, excludes a considerable part of words in any handout text.

In order to increase the coverage of the recognizer vocabulary, sentences can be modeled by n-grams of word segments, and the segments can be reconstructed into word forms after decoding. In order to evaluate this approach when applied to the Estonian language, two language models were created: a baseline system that uses word forms as language modeling units, and a system that uses subword units.

The language model training texts are processed by the morpheme analyzer, which marks the boundaries of word compounds in compound words, and marks the boundaries between stem and suffixes in inflected word forms. About 60% of the texts were also processed by a tool that expanded numbers and abbreviations into corresponding words. The tool also divided the text into sentences by looking at the punctuation marks and heuristically determining the sentence boundaries.

Many of the morpheme suffixes are very short (one phoneme suffixes are common). This increases the acoustic confusability of the units and decreases the span of the n-gram language model. To scope with this problem, we chose to decompose the words into stem and suffix only if the suffix is at least 2 phonemes long, in order to avoid the acoustically very confusable one-phoneme suffixes, although this causes the number of distinct language modeling units to grow and increases OOV rate. The suffixes are tagged so that the they are modeled separately from the stems that have the same orthography. The tagging also enables to reconstruct words forms from shorter segments after decoding.

After expanding the numbers and abbreviations, the training corpus contained 15 062 109 word forms. The number of words in the sentence was 13.7 on average. The number of distinct word forms was 742 194. The ratio of compound words in the distinct word list and in the corpus was 50.7% and 10.6%, respectively. After segmenting the corpus into pseudomorheme units, the number of tokens in the training text increased to 19 242 465. Average sentence length grew to 17.4. Total number of distinct words after splitting was 234 770, 351 (0.15%) of them were suffixes. The frequency of suffixes in the corpus was however much higher – 12.5%. The frequency of inflected words that were not split due to two phoneme constraint was 32.8% in the word list and 12.3% in the corpus.

The coverage of *N* most frequently occurring units was measured for vocabulary sizes from 10 000 up to 60 000, which is close to the maximum lexicon size of common decoders. Language modeling test set consist of the sentences in the BABEL speech database as well as two newspaper articles that were not present in the training corpus. The results are shown of figure 1. The out-of-vocabulary rate of 2.1%

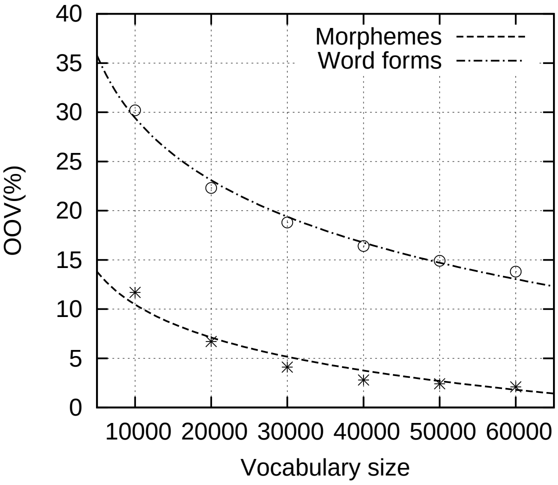

Figure 1: *Out-of-vocabulary rate for word-form and morpheme vocabularies.*

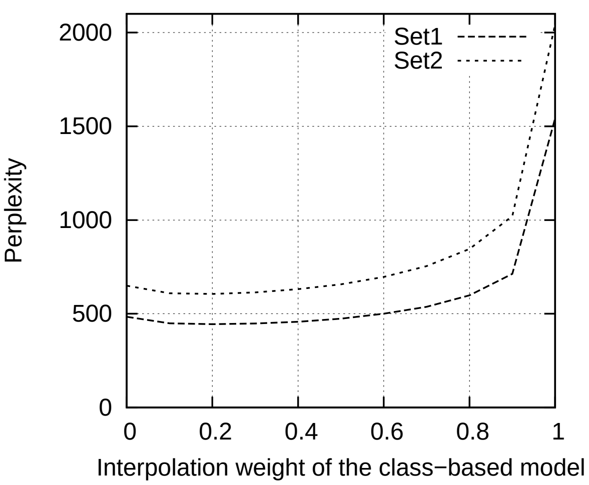

Figure 2: *Perplexity of the morpheme model when interpolated with the morpheme class model*

for a 60 000 unit dictionary of morphemes is similar with the results reported for other agglutinative languages [3, 2].

The lexicon of the language models was created from 60 000 most frequent units in the training data. The HTK language modeling tools were used for language model training. Good-Turing smoothing and a back-off value of 1 for both bigrams and trigrams was used. The language model performance was evaluated on two sets: sentences from two newspaper articles not present in the training data (“Set1”), and the sentences of the BABEL speech database that were later used for speech recognition experiments (“Set2”). The statistics of the word form and the morpheme models are given in table 1. Bigram and trigram hit values show the proportion of test set bigrams and trigrams found in the language models. The hit values for the two language models are similar for both test sets. The morpheme language model has a significantly better coverage and perplexity measure for the first test set, than the BABEL test set. This was expected because the models where trained mainly on newspaper texts, and there are some words and sentence constructs in the BABEL speech database that are not very common in everyday texts. The perplexity values for the morpheme models are much lower than those of the word form model, but they shouldn’t be compared with each other because the models operate on completely different token sets.

	Word forms		Morphemes	
	Set1	Set2	Set1	Set2
OOV rate	12.3%	14.4%	1.5%	2.9%
Bigram hits	66.3%	67.2%	75.6%	72.9%
Trigram hits	21.0%	23.5%	29.1%	27.5%
Perplexity	1448	1564	484	650

Table 1: *Language model statistics.*

The language models’ bigram and trigram hit ratios are relatively low for both sets. It is evident that due to the limited amount of training text and the free word order of the Estonian language, simple trigrams are not able to model word occurrence correlations with enough robustness. In order to better generalize to unseen and rare word sequences, a class-based morpheme trigram model was created. We used the word exchange algorithm [8] implemented in HTK to automatically derive 1000 morpheme classes from the training texts and trained a morpheme class trigram model. The bigram and trigram hit ratio of the class-based model is 94.2% and 53.3%, respectively. The perplexities of the class trigram interpolated with the morpheme trigram are shown of figure 2. The best perplexities were observed when the interpolation coefficient of the class-based model was 0.2 – the perplexities of the first and second test set were 444 and 606 (7-8% improvement over the pure morpheme model).

5. Recognition results

The recognition experiments were run using the Julius 3.4 multipath decoder [9]. For experiments that used interpolated language models, a modified version of Julius was used. The experiments were executed on a standard PC with an AMD Athlon 2600+ processor, 512 MB RAM, running Linux Mandrake.

We measured the word, morpheme and phoneme error rates of the recognition. Word error rates for the morpheme-based systems were acquired by reconstructing the words from recognized morpheme units after decoding. Phoneme error rates are measured by expanding the recognized units to their corresponding phonemes and comparing them with reference transcriptions. The results are shown in table 2.

We made some further experiments to find the optimal parameters for the interpolated system. We measured word error rates with different interpolation weights and different number of morpheme classes. It turned out that the system

Error rate	Word forms	Morphemes	Interpolated morphemes & classes
Words	37.7%	31.2%	28.3%
Morphemes	N/A	28.8%	26.6%
Phonemes	10.1%	8.0%	7.6%

Table 2: *Recognition results.*

performed best when the number of classes was 800 and weight of the class language model was 0.7 – the word error rate was then 27.7%. However, there was no big difference within the results, e.g when using the interpolation weight of 0.5, the difference between the worst system (with 600 classes) and the best system (with 800 classes) was only 1.32% absolute. Full results are shown on figure 3.

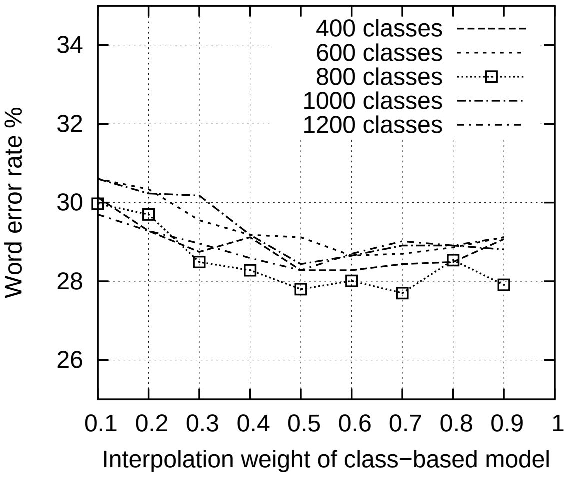

Figure 3: *Word error rate as a function of interpolation weight for models with different number of morpheme classes.*

6. Conclusions

Due to the large number of distinct word forms, traditional word based statistical language models for Estonian suffer from high out-of-vocabulary rates. To increase lexicon coverage, we used morphemes as basic units in the language model. In order to better model the probabilities of rare and unseen n-grams, the morphemes were clustered into classes and the class-based model was interpolated with the morpheme model. When applied to the speech recognition task, the resulting language model obtained a word error rate of 27.7%. This was significantly lower than word error rates when using a standalone morpheme model (31.2%) and the baseline word form model (37.7%).

7. Acknowledgments

This research was partly funded by the Estonian Information Technology Foundation as part of the Tiigriülikool program.

We thank Prof. Sadaoki Furui and Takahiro Shinozaki for assistance with the Julius decoder.

8. References

[1] A. Waibel, P. Geutner, L. Tomokiyo, T. Schultz, and M. Woszczyna, “Multilinguality in speech and spoken language systems”, Proceedings of the IEEE, vol. 88, no. 8, pp. 1297–1313, 2000.

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

[3] M. Szarvas and S. Furui, “Evaluation of the stochastic morphosyntactic language model on a one million word Hungarian dictation task”, Proceedings of Eurospeech, Geneva, 2003.

[4] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, “Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner”, Proceedings of Eurospeech, Geneva, 2003.

[5] A. Eek and E. Meister, “Estonian speech in the BABEL multi-language database: Phonetic-phonological problems revealed in the text corpus”, Proceedings of LP’98. Vol II., 1999, pp. 529–546.

[6] T. Hennoste, H.-J. Kaalep, K. Muischnek, L. Paldre, and T. Vaino, “The Tartu University corpus of Estonian literary language”, Abstracts. Congressus nonus internationalis fenno-ugristarum. Pars II, Tartu, 2000, pp. 338–339.

[7] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, “The HTK Book (for HTK Version 3.2),” http://htk.eng.cam.ac.uk, 2003.

[8] R. Kneser and H. Ney, “Improved clustering techniques for class-based statistical language modelling”, Proceedings of the European Conference on Speech Communication and Technology, 1993, pp. 973–976.

[9] A. Lee, T. Kawahara, and K. Shikano, “Julius — an open source real-time large vocabulary recognition engine”, Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH), 2001, pp. 1691–1694.