Initial Experiments with Estonian Speech Recognition

Anton Ragni
Department of Physics
University of Tartu
51010 Tartu, Estonia
ragni@ut.ee

Abstract

This paper presents a short description of work recently done at University of Tartu to construct a word–based speech recognition system. Simple bigram and trigram language models with cross–word triphone acoustic models are used by a one–pass best hypothesis recognizer to perform decoding of test data. The lowest word error rate of 37.5% reported in this paper is a common figure for word–based speech recognition of languages like Estonian.

1 Introduction

Estonian belongs to a family of inflectional and agglutinative languages which received a particular attention in recent years (Maučec et al., 2003; Kurimo et al., 2006). A single base word–form by means of inflections and compounding may have a huge number of derivative words. This greatly complicates the problem of building a speech recognition system with comparable word error rate (WER) performance to English systems. A common approach is to employ some type of subword systems, the goodness of which can be compared to each other and/or a word–based system (Hirsimäki et al., 2005). This paper is devoted to building of such word–based system and reports on results we obtained.

The first comprehensive description of work done on Estonian speech recognition appeared only recently (Alumäe, 2006). A huge number of experiments is conducted on two databases: Estonian part of Babel multi–language database (Eek and Meister, 1998) and Estonian SpeechDat–like database (Meister et al., 2003). The language modeling is performed both on a word and subword level. Our set of experiments is much more modest as compared to that work. However, we do not replicate the work already done but provide a completely independent set of results on Estonian part of Babel database.

The rest of the paper is organized as follows: in Section 2 language modeling is described. Section 3 is devoted to acoustic modeling. Section 4 describes experiments we performed and Section 5 makes conclusions drawn from this study.

2 Language Modeling

Experimental work with language modeling is performed on the Mixed Corpus of Estonian (MCE) – a set of written texts collected and maintained by University of Tartu[1] The total size of MCE is approximately 77M words excluding such special tags like sentence beginning (<s>), sentence ending (</s>) and number (<NUMBER>) symbols.[2]

A number of competitive bigram and trigram language models (LMs) is created using the HTK toolkit[3] The vocabulary of LMs is fixed to 65,000 most frequent words. All the diversity of LMs is obtained by application of different cut–off values to the number of bigrams and trigrams left in the model. The *cut–off value* specifies the least number of times any n–gram should have been seen in the

[1] http://www.cl.ut.ee/

[2] All numbers in this study are mapped to a common tag <NUMBER> since there is no known to us application capable of expanding them into verbal representations.

[3] http://htk.eng.cam.ac.uk/

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 249–252

corpus to be included in the model. Standard Good–Turing discounting is applied to refine parameters of LMs. The discounting factor k is kept greater from the cut–off value by seven for both bigrams and trigrams. Table 1 provides information about cut–off values for bigrams and trigrams, number of n–grams and size (for ARPA–compatible textual representation) of corresponding trigram LM. The size of trigram LM with cut–off values of 0 significantly exceeds the amount of available computer memory so it is not used in further experiments.

Cut-off	Bigrams	Trigrams	Size (MB)
0	11,676,757	34,166,450	–
1	3,855,881	5,760,565	111.7
10	635,555	507,714	14.3
100	66,440	38,885	2.4

Table 1: Characteristics of trigram LMs

3 Acoustic Modeling

3.1 Babel Multi–Language Speech Database

Experimental work with acoustic modeling is performed on Estonian part of Babel multi–language speech database (Eek and Meister, 1998). The database consists of three subsets: *very few*, *few* and *many* talker sets. The recordings are made in a clean recording environment from the set of 40 text passages, 2 sets of numbers and 4 sets of sentences with multiple occurrence of acoustically confusable words (e.g., *Lina* and *liina*, *türi* and *tüüri*). The recorded speech is sampled at 20,000 Hz and digitized using 16-bit integers.

The training part in this study is composed from the very few and many talker sets. The few talker set is used for development and testing. Basic statistics for training, development and testing parts is summarized in Table 2.

	Train	Dev	Test
Passages	163	40	40
Sentence groups	67	0	8
Number groups	64	0	8
Hours	7.4	0.3	0.9

Table 2: Statistics for training, development and testing parts of Babel Speech Database

All audio data in this study is preprocessed using Mel–Frequency Cepstral Coefficients (MFCC) feature extraction scheme with default values of configurable parameters (Young et al., 2006).

3.2 Unit Selection

The first step in acoustic modeling is to decide upon basic modeling units. There are many options to choose from: words, syllables, phonemes. The large vocabulary speech recognition is best done with phoneme units. There are two possible phoneme sets: orthographic and phonetic set. Experiments conducted on two different Estonian speech corpora revealed no preference in WER figures between these two representations (Alumäe, 2006). The orthographic representation used in this study is based upon the letters of Estonian alphabet with some minor modifications to the loaned letters such as *c*, *q*, *x*, etc. These letters are substituted with a sequence of common letters following the generic rules of Estonian pronunciation.

There are 32 letters in Estonian alphabet and 27 of them are considered to be common letters. The remaining 5 letters are substituted with one or more letters from the first set. In addition, two models are created for representing *short pause* (usually between two words) and *silence* (usually between two phrases or sentences) events. Thus the monophone set consists of 29 models:

```
a, b, d, e, f, g, h, i, j, k,
l, m, n, o, p, r, s, sh, z, zh,
t, u, v, io, ae, oe, ue, sp, sil
```

where `sh` corresponds to *š* letter, `zh` to *ž*, `io` to *õ*, `ae` to *ä*, `oe` to *ö* and `ue` to *ü*.

3.3 Acoustic Models

Acoustic model (AM) training follows the generic training procedure described in (Young et al., 2006): a single 3–state left–to–right hidden Markov model (HMM) is constructed for each monophone except for short–pause (`sp`) model which is a single state HMM tied to the central state of silence (`sil`) model; once the monophone models are trained, the next stage of training procedure is to create a set of cross–word triphone models the parameters of which are tied using a phonetic decision–tree state tying procedure implemented in HTK; the number

of mixtures is gradually increased until each model in the final set is represented at each state by the weighted sum of eight Gaussian probability density functions (pdfs).

4 Experiments

A large vocabulary speech recognizer implemented in the HTK toolkit (`HDecode`) is used to transcribe test sentences. There are many configurable parameters to alter the decoding process in some direction (speed, depth, accuracy, etc). For some of them the default values are used, for others, values giving the lowest WER are estimated on the development set.

Fig. 1 gives an example of influence imposed by the value of *word insertion penalty* on resulting WER figures. The word insertion penalty is a fixed

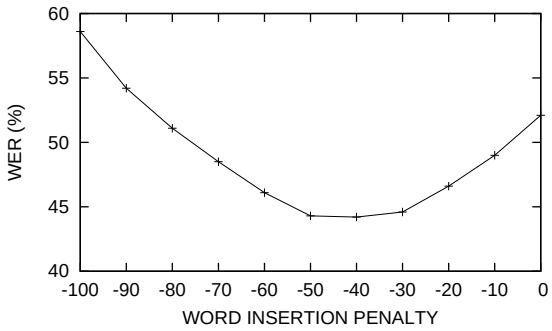

Figure 1: WER at different values of word insertion penalty

value added to each token when it transits from the end of one word to the beginning of another. By penalizing inter–word tokens we can force the appearance of new words only when their probability becomes sufficiently high.

Fig. 2 shows the performance of recognizer at different widths of decoding beam. The *main beam width* restricts the growth of recognition network and token propagation only to those HMM models the likelihoods of which fall no more than a beam width from the most likely model. Narrow beam width results in a smaller number of tokens considered at any given time, thus increasing the decoding speed. The time spent for recognition of test utterance is usually given as a portion of utterance real length called real–time ratio (RT). In Fig. 2 the decoding time varies between 0.2xRT and 7xRT. (Estimation of parameters and evaluation of test data is performed at the main beam width value of 200 with 2xRT speed of decoding unless otherwise stated.)

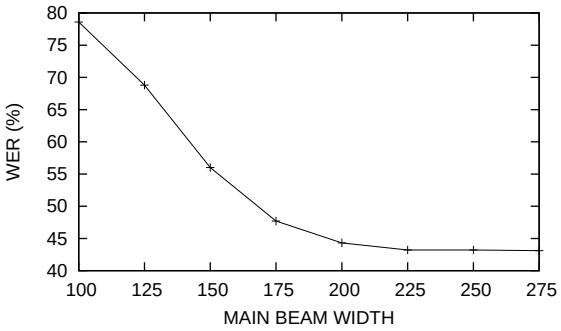

Figure 2: WER at different values of main beam width

Fig. 3 shows WER figures on the development set when LM and AM likelihoods are scaled. Giving a

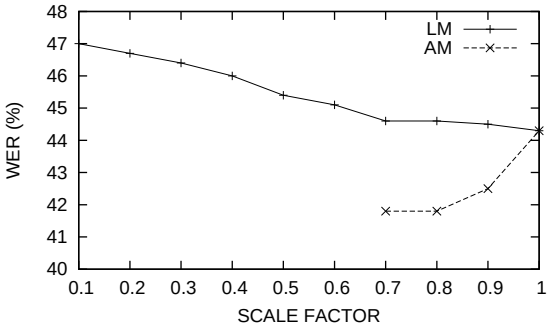

Figure 3: Effect of scaling down AM and LM likelihoods (results for up to 10xRT performance are shown)

preference to AM likelihoods by scaling down LM likelihoods leads to increased error rates; decreasing AM likelihoods, on contrary, enhances the accuracy of recognition with degradation in RT–performance.

Fig. 4 shows the improvement in recognition accuracy obtained by incrementing number of pdfs in the HMM state output distribution. The major drop in WER occurs when the number of pdfs is increased from three to five (13.4% absolute or 25.0% relative). Additional pdfs, however, lead to negligible reduction in WERs (1.8% absolute or 4.5% relative).

Final results of evaluation are given in Table 3. The first column describes the LM used in recog-

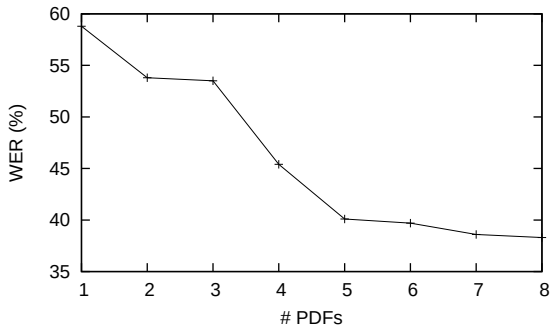

Figure 4: WER at different number of pdfs in HMM state distribution

LM	Del.	Subs.	Ins.	WER
bg100	420	1494	89	38.9
bg10	417	1472	86	38.3
bg1	417	1447	84	37.8
bg0	416	1438	81	37.5

LM	Del.	Subs.	Ins.	WER
tg100	420	1494	89	38.9
tg10	417	1471	86	38.3
tg1	417	1446	84	37.8

Table 3: Number of errors and corresponding WER figures for bigram and trigram LMs

nition: `bg` stands for bigram, `tg` for trigram and appended with the cut–off value for any order of n–grams. Next three columns specify the amount of different errors made by recognizer. As can be noted, 75% of all errors are substitution errors which can be originated from the lack of proper n-grams or weakly representative AMs, or both. LM's hit–ratios on the testing set confirm at least the first assumption: average hit–ratios for bigrams and trigrams are 50% and 11%. Since the number of trigram hits is very low, in major cases bigram instances are used instead. This also explains a comparable WER figures of bigram and trigram LMs. The lowest WER of 37.5% is obtained by the most comprehensive LM – bigram LM with more than 11M of distinct bigrams (113MB).

5 Conclusions

In this paper we described briefly the initial set of experiments with Estonian speech recognition using Estonian part of Babel speech database. The lowest WER reported in this paper (37.5%) can be compared to recently reported value of 36.2% (Alumäe, 2006) if we account for reduced amount of training data available to us. The recognition of test data is performed using a single–pass best hypothesis strategy which generally loses considerably to multi–pass N–best list strategies with a lattice stage rescored using more comprehensive LMs. However, this is an example of things needed to be done in the future using baseline systems built and described in this paper.

References

T. Alumäe. 2006. *Methods for Estonian Large Vocabulary Speech Recognition*. PhD thesis, Tallinn University of Technology.

A. Eek and E. Meister. 1998. Estonian speech in the babel multi–language database: Phonetic–phonological problems revealed in the text corpus. In *LP*, volume II, pages 529–546.

T. Hirsimäki, M. Creutz, V. Siivola, and M. Kurimo. 2005. Morphologically motivated language models in speech recognition. In *AKRR*, pages 121–126, Espoo, Finland.

M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *HLT-NAACL*, New York, USA.

M. Maučec, T. Rotovnik, and M. Zemljak. 2003. Modelling highly inflected slovenian language. *Int. J. of Speech Tech.*, 6:245–257.

E. Meister, J. Lasn, and L. Meister. 2003. Development of the Estonian SpeechDat–like database. In *Proceedings of Eurospeech*, pages 1601–1604, Geneva, Switzerland.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2006. *The HTK Book*. Cambridge University Press.