Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)
© 2012 The Authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-133-5-10

Transcription System for Semi-Spontaneous Estonian Speech

Tanel ALUMÄE [1]
Institute of Cybernetics at Tallinn University of Technology, Estonia

Abstract. This paper describes a speech-to-text system for semi-spontaneous Estonian speech. The system is trained on about 100 hours of manually transcribed speech and a 300M word text corpus. Compound words are split before building the language model and reconstructed from recognizer output using a hidden event *N*-gram model. We use a three pass transcription strategy with unsupervised speaker adaptation between individual passes. The system achieves a word error rate of 34.6% on conference speeches and 25.6% on radio talk shows.

Keywords. Estonian, speech recognition, compound words

Introduction

This paper describes the current offline speech-to-text transcription system for semi-spontaneous speech for the Estonian language.

Estonian is the official language of Estonia, spoken by about one million people. The need for support for language technology has been recognized by the Estonian government and the field is now actively supported. In the context of the national program Estonian Language Technology 2011-2017 and its predecessor [1], several applications that use offline large vocabulary continuous speech recognition (LVCSR) for Estonian have been developed at the Laboratory of Phonetics and Speech Technology at the Institute of Cybernetics at Tallinn University of Technology. The Transcribed Speech Archive Browser [2] is a web application that provides access to hundreds of hours of automatically transcribed radio broadcasts (mainly conversational programs and long news broadcasts). End-users can use a web browser to navigate in the hierarchy speech recordings, search from the transcriptions, view the transcriptions and listen to the recordings. Another application targetted towards end-users is the web-based speech transcription service[2]. The service allows users to transcribe long speech files by uploading them to the lab's server. The speech files are transcribed on the server and the resulting transcripts are sent back via e-mail. The service can also be used through a simple API. Third application, *Diktofon*[3], is an application for the Android smart-phone platform that provides basic digital voice recorder functionalities. In addition, it uses the previously described

[1] Corresponding author. E-mail: tanel.alumae@gmail.com.
[2] `http://bark.phon.ioc.ee/webtrans/`
[3] `http://play.google.com/store/apps/details?id=kaljurand_at_gmail_dot_com.diktofon`

web-based speech transcription service to provide automatic transcripts for recordings containing Estonian speech.

This paper gives a detailed description of the multi-pass transcription system that serves the applications. We describe the corpora and methods used for training the acoustic and language models, outline the multi-pass transcription strategy and compare two methods for reconstructing compound words from the system output. Transcription quality is reported on two domains: conference speeches and broadcast conversational speech.

Estonian speech recognition has been studied before in several publications. As Estonian is a language with a fairly complex morphology, most experiments have targeted word decomposition for language modeling (e.g., [3]).

1. System Description

1.1. Speech Data

For training the acoustic models (AMs), we used various wideband Estonian speech corpora, totalling in about 97 hours:

- the BABEL speech database [4] which contains about 9 hours of dictated speech;
- a corpus of Estonian broadcast news which contains mostly dictated speech, with some semi-spontaneous studio and telephone interviews (16 hours);
- a corpus of broadcast conversations which consists of various talk shows from three radio stations, mostly discussing political matters in a semi-spontaneous studio setting (20 hours);
- a corpus of semi-spontaneous telephone interviews from radio news programs, discussing mainly daily news and current events (18 hours);
- a corpus of local conference speeches, recorded with a close-talking microphone (18 hours);
- a corpus of studio-recorded spontaneous monologues and dialogues [5] (16 hours)

For tuning and measuring system performance, two different domains were used: conference speeches and broadcast conversations, with separate development and test sets for both of the domains. The conference domain development and test sets both consist of three 20-minute presentation recordings from a local linguistic conference. The development set for the broadcast conversations domain contains four different radio talk shows from 2009, each about 45 minutes, with 11 unique speakers in total. For testing, two different sets were used: seven radio talk shows from 2011, each about 45 minutes (17 speakers in total), and a set of 10 broadcast telephone interviews from 2011 (41 minutes, about 20 speakers). None of the development and test data was used for training.

1.2. Acoustic Models

The RWTH ASR toolkit [6] was used for training AMs. The models are continuous triphone HMMs with 2000 Gaussian mixtures that use 385150 Gaussian distributions. Decision-tree based clustering, with manually defined phonetic questions, was used to

Table 1. Phonemes defined in the acoustic model, with corresponding IPA symbols, and example words with their pronunciations as defined in the system pronunciation lexicon.

Vowels			Consonants		
Phoneme	IPA	Examples	Phoneme	IPA	Example
a	a	kalu /k a l u/, kaalu /k a a l u/	k	g̊	lagi /l a k i/, üheksa /ü h e k s a/
e	e	elu /e l u/	p	b̥	kabi /k a p i/
i	i	ilu /i l u/	t	d̥, d̥ʲ	padu /p a t u/, padi /p a t i/
o	o	kole /k o l e/	k:	k	laki, lakki /l a k: i/
u	u	usin /u s i n/	p:	p	kapi, kappi /k a p: i/
õ	ɤ	õlu /õ l u/	t:	t, tʲ	patu, pattu /p a t: u/
ä	æ	kära /k ä r a/	l	l, lʲ	kallas /k a l l a s/
ö	ø	kört /k ö r t:/	r	r	nari /n a r i/
ü	y	tühi /t ü h i/	m	m	samu /s a m u/
			n	n, nʲ	hani /h a n i/
			v	v	kava /k a v a/
Non-speech units			f	f	foori /f o o r i/
Silence/filler		Silence, breathing, hesitation, etc	j	j	m a j a /m a j a/, majja /m a i j a/
Garbage		Unintelligible speech	h	h	sahin /s a h i n/
			s	s, sʲ	kassi /k a s s i/
			š	ʃ	tuši /t u š i/, garaaž /k a r a a š/

create tied-state cross-word triphones. Maximum likelihood training was used for creating the models.

The AM inventory contains 25 phoneme models, a silence/noise model and a garbage model that is used to absorb unintelligible and foreign language words during training (see Table 1). Although different noises and fillers are annotated in our training data at a relatively fine-grained level, they are all mapped to a single silence/noise model during training. We also merge palatalized and unpalatalized versions of the various phonemes into single acoustic units, since it is difficult to derive the correct palatalization from the orthographic words forms. Long phonemes are modeled using a sequence of two short phoneme units, except for plosives which have dedicated models for long versions. The overlong quantity degree, a prominent factor of the Estonian phonetic system, is not modeled at all, since it is difficult to model the quantity degrees using segmental (phoneme) units. The other reason behind such simplifications is the fact that distinction between palatalized and unpalatalized phonemes, as well as distinction between the long and overlong quantity degrees, is not usually needed for discrimination between orthographic word forms (i.e., palatalization and overlong quantity (with some exceptions) is not visible in orthography).

Each 45-dimensional acoustic feature vector is calculated by applying linear discriminant analysis (LDA) to the concatenation of nine neighboring 16-dimensional feature vectors representing Mel-frequency cepstral coefficients (MFCCs), calculated from the 16 kHz audio signal using a 10 ms frame shift and a 25 ms Hamming window. Cepstral mean normalization over individual speech segments is applied.

Table 2. Language model training data

Source	Documents	Tokens	Source	Documents	Tokens
Newspapers	655 847	206M	Fiction	202	6.3M
Web news portals	186 781	40M	Broadcast conversations	227	0.34M
Scientific publications	78 709	17M	Blogs	3722	0.17M
Parliament transcripts	6024	15M	Conference transcripts	23	0.06M
Magazines	4137	12M			
			Total	935 672	299M

1.3. Language Model

Text data sources used for training the language models (LMs) are listed in Table 2. Most of the written language corpora are compiled at the University of Tartu [7]. In order to have up-to-date language data, we scraped additional web data from news portals and blogs. Finally, transcriptions of conversational broadcast data (talk shows, telephone interviews) and conference speeches were used as a sample for spoken language.

Before using the text data for LM training, text normalization is performed. Raw texts are processed by the morphological analyzer [8] which splits text into sentences, tokenizes, recapitalizes, assignes word types and morphological attributes to words and annotates words with morphological structure. Numbers are expanded into words; in Estonian, number expansion depends on the inflection of the numbers which is usually not lexicalized in orthographic form; instead, morphological attributes of the surrounding words can be used to disambiguate between different inflectional forms of the numbers; we used a small set of hand-written rules for disambiguation, although a statistical classifier would probably work better. Abbreviations, such as `jne` (Eng `etc`), `km`, are expanded into words, using a hand-built mapping table.

As Estonian is a heavily compounding and inflective language, the lexical variety of the language is very high [9]. To reduce the out-of-vocabulary (OOV) rate of the LM, compound words are decomposed into compound particles, using the word structure information assigned by the morphological analyzer. The result of such splitting on the OOV rate is depicted on Figure 1. Compound splitting gives especially drastic results in the conference speech domain, where speech contains many scientific terms that are often compounds: even with a 1M word lexicon, the OOV-rate of a full word vocabulary is almost 5%, while when using compound splitting, the OOV rate is below 3% with a 200K vocabulary. Although in many previous experiments ([3,9]), morphemes have been used as basic units in Estonian LVCSR, we have found that decomposing only compound words and using a larger lexicon instead helps to avoid acoustic confusability of short morphemes, improves the span of a 4-gram LM and avoids outputting ungrammatical words as a result of concatenating illegal morpheme sequences.

LM vocabulary is created by selecting the 200 000 most likely case-sensitive compound-split units from the mixture of the corpora, given the development texts. For each corpus, a 4-gram LM is built. The LMs are compiled by including all bigrams and trigrams as well as 4-grams occurring more than once, using interpolated modified Kneser-Ney discounting. The individual LMs are interpolated into one by using interpolation weights optimized on development data. Finally, the LM is pruned to about one third in size using entropy pruning. This is all done using the SRILM toolkit [10].

Table 3. Out-of-vocabulary rates and language model perplexities.

	Conference speech LM		Broadcast conversation LM		
	Conference speeches		Radio talk shows		Telephone interviews
	Dev	Test	Dev 2009	Test 2011	Test
OOV	2.9%	3.0%	1.5%	0.7%	0.7%
Perplexity	603	644	435	370	390

For the current work, two LMs were built: for conference speeches and for conversational broadcast speech. For both domains, the LM was optimized on the development data of the corresponding domain (see section 1.1). The out-of-vocabulary (OOV) and perplexity results of the LMs are listed in Table 3.

1.4. Pronunciation Lexicon

As Estonian is a language with a close relationship between word orthography and pronunciation (see Table 1), a rule based system was used for deriving the pronunciations for words in the LM lexicon [4]. The rules are mainly concerned with determining the correct variant (short or (over)long) of the plosives (/k/, /p/, /t/) based on the usage context. In addition, for some more common words and word types, additional reduced pronunciation variants are assigned: e.g., for the -nud words, such as `saanud` (Eng `got`), a reduced pronunciation (/s a a n t/) was assigned in addition to the canonical form (/s a a n u t/).

For about 200 most common foreign proper names and abbreviations, pronunciation was assigned by hand.

1.5. Decoding Strategy

The transcription system consists of a speech detection and speaker diarization module, followed by a three-pass decoding module, and the final compound word reconstruction module.

[4] Available at `https://github.com/alumae/et-pocketsphinx-tutorial/blob/master/scripts/est-l2p.py`

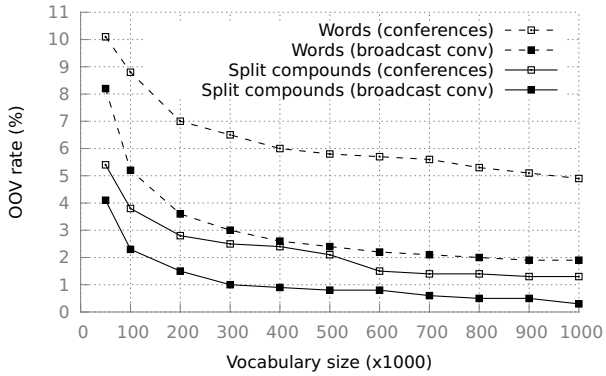

Figure 1. OOV-rates for two development sets with various vocabulary sizes, when using full words or compound-split words as basic units.

Input audio data is segmented into shorter sentence like chunks using the LIUM SpkDiarization [11] toolkit. Segments are classified as speech or non-speech using a Gaussian mixture model built from our AM training data. Segments containing speech are clustered, with each cluster corresponding ideally to one unique speaker in the recording. BIC clustering followed by CLR-like clustering [12] are applied. The resulting speaker labels are used to perform unsupervised AM adaptation in the multi-pass decoding system.

In the first decoding pass, speaker-independent AMs are used. The resulting hypotheses are used for creating CMLLR adaptation [13] transforms for each speaker which are then used in the second pass. The second pass hypotheses are similarly used for MLLR adaptation [14] (based on CMLLR-transformed features). The last pass which combines both CMLLR and MLLR adaptation results in a word lattice for each segment. Each of the three passes runs in roughly two times slower than realtime. The resulting lattices are decoded using consensus decoding [15].

1.6. Reconstructing Compound Words

Output of the decoding passes consists of sequences of word-like units where compound words (including words containing a hyphen as separator) are replaced with their compound constituents. In the final system output, compound words have to be reconstructed from the constituents.

One method for compound reconstruction was proposed in [9]. This approach relies on an *N*-gram LM which is trained on a text corpus where compound words are segmented into their components, with special non-word tokens (”connector tags”) between the constituents. The compound word reconstruction problem can then be treated as a problem of recovering the hidden connector tags between the words. The latter can be solved by decoding the input sequence using the Viterbi algorithm which finds the most likely sequence of words and hidden tokens, based on the trained hidden event LM. Finally, compound words are reconstructed in places where the hidden connector tags were inserted between words. In this work, we slightly extended this approach by modeling a second hidden token for the dash-connected compound words (as in `võib-olla`, Eng lit `may-be` and `Lõuna-Eesti`, Eng lit `South-Estonia`).

We considered an alternative method for reconstructing compound words based on conditional random fields (CRFs). We used the CRF model to classify the tokens into three classes: ”intra-compound” (i.e., the token should be concatenated with the next token), ”hyphen-compound” (i.e., the token should be concatenated with the next token using a hyphen) and ”not-compound”. One of the benefits of CRFs over the hidden event LM is that they are based on features rather than only observed words, i.e., the probability calculations are not only based on word identities but also on other features such word prefixes and suffixes. We used the following features in the CRF model: five unigram features (word itself, two previous and two next words), two bigram features (previous and current word, current word and next word), unigram features that depend on the suffixes of previous, current and next word, word capitalization features of the previous, current and next words and current word prefix features. The CRF model is trained based on a random sample (about one fifth) of LM training data since otherwise training exhausted the 32 GB memory available on the system. We also experimented with using more training data with a poorer feature set but it didn’t improve the accuracy. We used the Wapiti toolkit [16] in all experiments.

Table 4. Compound tag insertion accuracies with hidden-event LM and CRF model. Last column shows the WER of a transcription task when using the corresponding model.

Model	Tag	Precision	Recall	F1	WER
Hidden event LM	Compound	0.97	0.89	0.93	25.0%
	Dash	0.85	0.44	0.58	
CRF	Compound	0.94	0.87	0.90	25.2%
	Dash	0.83	0.33	0.48	

Table 4 lists precision, recall, and F1 measure for finding the two hidden tags when using the two different models. Although the CRF model uses a richer feature set, it doesn't achieve quite the same accuracies as the hidden event LM, although the difference has only a very small impact on the final transcription WER. In the reported speech transcription experiments, we used the better-performing hidden event LM for reconstructing compound words.

2. Experimental Results

Word error rate results for all the sets are listen in Table 5. For both domains, the LM weight was tuned on the development set. No other hyperparameter optimization was performed.

The highest WER is observed on the conference speech data sets. This could be expected, since this domain has also higher OOV rate and LM perplexity values, compared to the broadcast conversations domain. Perhaps surprisingly, the WER of telephone interviews is almost the same as that of the studio-recorded broadcast conversations, although telephone interviews have a narrower audio bandwidth and feature more spontaneous speech.

The multi-pass transcription strategy with consensus decoding achieves 3-7% absolute (11-18% relative) WER reduction, compared to the speaker independent approach.

3. Conclusion

The paper described a LVCSR system for Estonian semi-spontaneous speech. The system performance was measured in two domains: speeches of a local linguistic conference and broadcast conversations in local radio stations. The system uses a 4-gram LM with a 200K vocabulary where compound words are decomposed into compound particles. The latter helps to decrease the lexicon OOV rate from 7% to 3% for the conference domain and from 5% to 1.5% for the broadcast conversations domain. Compound words are

Table 5. Final system word error rates (%) for various sets after each decoding step.

	Conference speeches		Radio talk shows		Telephone interviews
Step	Dev	Test	Dev 2009	Test 2011	Test
Speaker independent	38.5	38.8	28.1	29.5	32.0
+CMLLR	34.9	37.2	26.1	27.7	28.9
+MLLR	32.2	35.3	24.9	26.2	27.1
+CN	31.5	34.6	24.9	25.6	26.6

reconstructed from system output using a hidden event LM that is found to outperform CRF models. The system applies unsupervised AM adaptation using CMLLR and MLLR transforms which result in WERs of 34.6% and 26% for the respective domains.

Future work includes experimenting with discriminative and speaker-adaptive training for AMs, modeling spontaneous speech phenomena in the pronunciation model and using more advanced LM adaptation techniques.

Acknowledgments

This research was partly funded by the Estonian Ministry of Education and Research target-financed research theme no. 0140007s12.

References

[1] E. Meister, J. Vilo, and N. Kahusk, "National programme for Estonian language technology: a pre-final summary," in *Baltic HLT 2010*, 2010, pp. 11–14.

[2] T. Alumäe and A. Kitsik, "TSAB – web interface for transcribed speech collections," in *Interspeech 2011*, Florence, Italy, 2011, pp. 3335–3336.

[3] E. Arısoy, M. Kurimo, M. Saraclar, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and H. Sak, *Speech Recognition, Technologies and Applications*. I-Tech, 2008, ch. Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages, p. 550.

[4] A. Eek and E. Meister, "Estonian speech in the BABEL multi-language database: Phonetic-phonological problems revealed in the text corpus," in *Proceedings of LP'98*, vol. 2, Prague, Czech Rep., 1999, pp. 529–546.

[5] P. Lippus, "The acoustic features and perception of the estonian quantity system," Ph.D. dissertation, Tartu University, 2011.

[6] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, "The RWTH Aachen University open source speech recognition system," in *Interspeech 2009*, Brighton, U.K., 2009, pp. 2111–2114.

[7] H.-J. Kaalep and K. Muischnek, "The corpora of Estonian at the University of Tartu: the current situation," in *Baltic HLT 2005*, Tallinn, Estonia, 2005, pp. 267–272.

[8] H.-J. Kaalep and T. Vaino, "Complete morphological analysis in the linguist's toolbox," in *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, Tartu, Estonia, 2001, pp. 9–16.

[9] T. Alumäe, "Automatic compound word reconstruction for speech recognition of compounding languages," in *Proceedings of NODALIDA*, Tartu, Estonia, 2007, pp. 5–12.

[10] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of ICSLP*, vol. 2, Denver, USA, 2002.

[11] S. Meignier and T. Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, Dallas, TX, USA, 2010.

[12] C. Barras, X. Zhu, S. Meignier, and J. L. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1505–1512, Aug. 2006.

[13] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75 – 98, 1998.

[14] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

[15] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373 – 400, 2000.

[16] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale CRFs," in *Proceedings of ACL*, Uppsala, Sweden, July 2010, pp. 504–513.

Human Language Technologies
The Baltic Perspective

Proceedings of the Fifth International Conference Baltic HLT 2012

Edited by

Arvi Tavast

Institute of the Estonian Language

Kadri Muischnek

University of Tartu

and

Mare Koit

University of Tartu

IOS Press

Amsterdam • Berlin • Tokyo • Washington, DC

© 2012 The Authors and IOS Press.

All rights reserved. No part of this book may be reproduced, stored in a retrieval system,
or transmitted, in any form or by any means, without prior written permission from the publisher.

ISBN 978-1-61499-132-8 (print)
ISBN 978-1-61499-133-5 (online)
Library of Congress Control Number: 2012947888

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

Distributor in the USA and Canada
IOS Press, Inc.
4502 Rachael Manor Drive
Fairfax, VA 22032
USA
fax: +1 703 323 3668
e-mail: iosbooks@iospress.com

LEGAL NOTICE
The publisher is not responsible for the use which might be made of the following information.

PRINTED IN THE NETHERLANDS

Frontiers in Artificial Intelligence and Applications

FAIA covers all aspects of theoretical and applied artificial intelligence research in the form of monographs, doctoral dissertations, textbooks, handbooks and proceedings volumes. The FAIA series contains several sub-series, including "Information Modelling and Knowledge Bases" and "Knowledge-Based Intelligent Engineering Systems". It also includes the biennial ECAI, the European Conference on Artificial Intelligence, proceedings volumes, and other ECCAI – the European Coordinating Committee on Artificial Intelligence – sponsored publications. An editorial panel of internationally well-known scholars is appointed to provide a high quality selection.

Series Editors:
J. Breuker, N. Guarino, J.N. Kok, J. Liu, R. López de Mántaras,
R. Mizoguchi, M. Musen, S.K. Pal and N. Zhong

Volume 247

Recently published in this series

Vol. 246. H. Fujita and R. Revetria (Eds.), New Trends in Software Methodologies, Tools and Techniques – Proceedings of the Eleventh SoMeT_12
Vol. 245. B. Verheij, S. Szeider and S. Woltran (Eds.), Computational Models of Argument – Proceedings of COMMA 2012
Vol. 244. S. Scheider, Grounding Geographic Information in Perceptual Operations
Vol. 243. M. Graña, C. Toro, J. Posada, R.J. Howlett and L.C. Jain (Eds.), Advances in Knowledge-Based and Intelligent Information and Engineering Systems
Vol. 242. L. De Raedt, C. Bessiere, D. Dubois, P. Doherty, P. Frasconi, F. Heintz and P. Lucas (Eds.), ECAI 2012 – 20th European Conference on Artificial Intelligence
Vol. 241. K. Kersting and M. Toussaint (Eds.), STAIRS 2012 – Proceedings of the Sixth Starting AI Researchers' Symposium
Vol. 240. M. Virvou and S. Matsuura (Eds.), Knowledge-Based Software Engineering – Proceedings of the Tenth Joint Conference on Knowledge-Based Software Engineering
Vol. 239. M. Donnelly and G. Guizzardi (Eds.), Formal Ontology in Information Systems – Proceedings of the Seventh International Conference (FOIS 2012)
Vol. 238. A. Respício and F. Burstein (Eds.), Fusing Decision Support Systems into the Fabric of the Context
Vol. 237. J. Henno, Y. Kiyoki, T. Tokuda, H. Jaakkola and N. Yoshida (Eds.), Information Modelling and Knowledge Bases XXIII
Vol. 236. M.A. Biasiotti and S. Faro (Eds.), From Information to Knowledge – Online Access to Legal Information: Methodologies, Trends and Perspectives

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)