Transcription System for Semi-Spontaneous Estonian Speech

Tanel Alumäe

Institute of Cybernetics at Tallinn University of Technology

Baltic HLT 2012, Tartu, 2012-10-05

Contents

Background

- Transcription system for semi-spontaneous Estonian speech
- Already used in several places:
  - Browse transcribed radio programs: `http://bark.phon.ioc.ee/tsab`
  - Web service for transcribing user-provided audio content: `http://bark.phon.ioc.ee/webtrans/`
  - Transcribing voice-recorder for Android (*Diktofon*)
  - Commercial interest from media monitoring companies

Dev and test sets

Transcription quality was measured in two domains:

1. Speeches of a local linguistic conference
   - Dev: 3 speeches
   - Test: 3 speeches, all 20 minutes
2. Broadcast conversations
   - Dev: 4 talk show from 2009, all 45 minutes, 11 speakers
   - Test 1: 7 talk shows from 2011 (17 speakers)
   - Test 2: 10 radio interviews from 2011 (41 minutes, 20 speakers)

Acoustic models

Training data

Corpus	**Type**	**Size**
BABEL speech database	dictated	8 h
Corpus of broadcast news	mostly dictated	16 h
Corpus of broadcast conversations (discussion programs)	semi-spontaneous	20 h
Corpus of telephone interviews from radio news programs	semi-spontaneous	18 h
Corpus of local conference speeches	partly semi-spontaneous	18 h
Corpus of studio-recorded spontaneous monologues and dialogues	spontaneous	16 h
Total		**97h**

Acoustic models

Inventory

Vowels		
Phoneme	IPA	Examples
a	ɑ	kalu /k a l u/, kaalu /k a a l u/
e	e	elu /e l u/
i	i	ilu /i l u/
o	o	kole /k o l e/
u	u	usin /u s i n/
õ	ɤ	õlu /õ l u/
ä	æ	kära /k ä r a/
ö	ø	kört /k ö r t:/
ü	y	tühi /t ü h i/

Non-speech units		
Silence/filler		Silence, breathing, hesitation, etc
Garbage		Unintelligible speech

Consonants		
Phoneme	IPA	Example
k	g̥	lagi /l a k i/, üheksa /ü h e k s a/
p	b̥	kabi /k a p i/
t	d̥, d̥^{j}	padu /p a t u/, padi /p a t i/
k:	k	laki, lakki /l a k: i/
p:	p	kapi, kappi /k a p: i/
t:	t, t^j	patu, pattu /p a t: u/
l	l, l^j	kallas /k a l l a s/
r	r	nari /n a r i/
m	m	samu /s a m u/
n	n, n^j	hani /h a n i/
v	v	kava /k a v a/
f	f	foori /f o o r i/
j	j	m a j a /m a j a/, majja /m a i j a/
h	h	sahin /s a h i n/
s	s, s^j	kassi /k a s s i/
š	ʃ	tuši /t u š i/, garaaž /k a r a a š/

Acoustic models

Details

- 25 phonemes + 2 non-phoneme sounds
- Relatively few individual sounds
  - Palatalized and unpalatalized phonemes merged
  - Long duration represented using a sequence of two short sound models (except for plosives)
  - Inter-word silence and filler sounds (breathing, hesitation, lip-smack, etc) merged into one
- Unintelligible speech and foreign words in training modeled using the garbage model

Technical

- we use the RWTH-ASR toolkit (open source, free for non-commecial use)
- 9 MFCC frames merged by LDA to 45-dim feature vector
- Continuous triphone HMMs, 2000 Gaussian mixtures, 385 000 Gaussians, decision-tree based triphone clustering

Language model

Training data

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

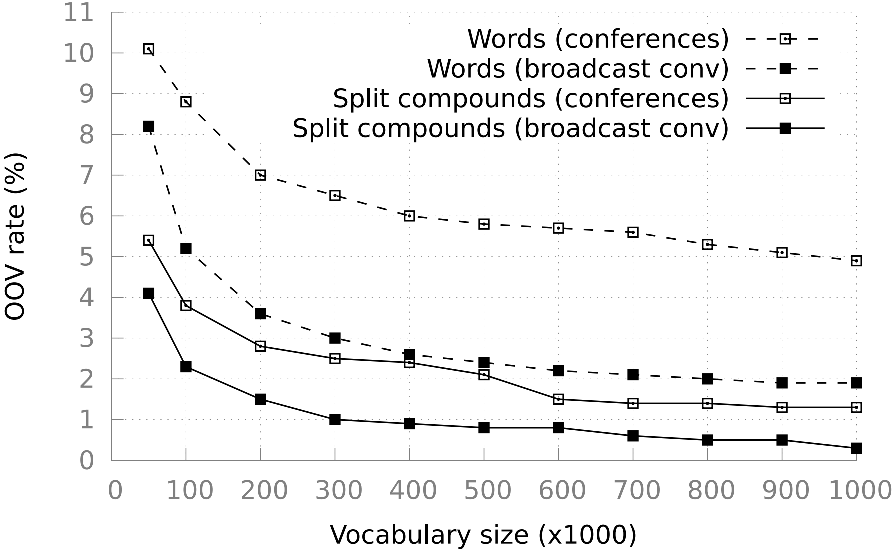
Language model

Text normalization

1. Process texts using a morphological analyzer (Filosoft)
   - splits texts into sentences
   - tokenizes
   - recapitalizes
   - assigns morphological attributes to words
   - annotates words with morphological structure
2. Expand numbers into words, inflection guessed from context
3. Expand abbreviations
4. Split compound words

Language model

Words *vs* compound-split words



Language model

Training

- 200K ”word” vocabulary, case-sensitive
- 4-gram LM from each of the corpora, interpolated into one
- Finally pruned to 1/3 in size

Pronunciation lexicon:

- Hand-written G2P rules (very simple)
- About 200 exceptions for common foreign names

Model details

	Conference speech LM	Broadcast conversation LM	
	Conference speeches	Radio talk shows	Teleph. interviews
OOV rate	3.0%	0.7%	0.7%
Perplexity	644	370	390

Decoding

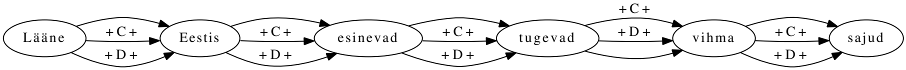
1. Segmentation of audio into sentence-like chunks
2. Speech/non-speech classification
3. Segment clustering according to speaker (speaker diarization)
4. Decoding using speaker-independent models
5. CMLLR adaptation, re-decoding
6. MLLR adaptation, re-decoding into word lattice
7. Confusion decoding of word lattice
8. Compound word reconstruction

Compound word reconstruction

Problem

Input: *Lääne Eestis esinevad tugevad vihma sajud*
Goal: ***Lääne-Eestis*** *esinevad tugevad* ***vihmasajud***

Solution: use hidden-event language model, find the most probable path, using a trigram language model that has extra hidden units for inter-word dash and inter-word "compound break" marker.

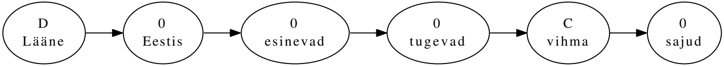


Output: *Lääne +D+ Eestis esinevad tugevad vihma +C+ sajud*

Compound word reconstruction: alternative method

Alternative method: conditional random fields (CRF)

- Treat as a sequence labeling problem
- Often used for tasks such as Named Entity Recognition
- Label each word as ”simple” (0), ”append-next” (C), ”dash-next” (D)
- Look at features of the word and its neigbours: word itself, prefix, suffix, shape, etc
- Feature weights learned from data using gradient descent
- Decoding gives the most likely label sequence given the input
- However: training requires more memory, cannot use all data



Compound word recognition: results

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	

Transcription results

Word error rate

	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

- multi-pass transcription strategy with consensus decoding achieves 3-7% absolute (11-18% relative) WER reduction

Future work

- Goal: reduce WER to 20%
  - More training data
  - Discriminative training
  - Unsupervised techniques
  - More advanced acoustic features