Conversational Telephone Speech Recognition for Lithuanian

Rasa Lileikyte, Lori Lamel, Jean-Luc Gauvain
LIMSI-CNRS

SLSP, 2015

Introduction

Goal: develop a conversational telephone speech recognition system for the low-resourced Lithuanian language

Questions

- Phoneme-based system better than grapheme-based system?
- How can additional resources improve system performance?
  1. Web texts
  2. Untranscribed audio

Outline

- Conversational telephone speech
- Description of Lithuanian language
- Speech-to-text and Keyword spotting systems
- Data set
- Baseline recognition systems
- Experiments
- Conclusions

Conversational Telephone Speech

Transcribing conversational telephone speech is a complex task

- Silence can be inserted between words
- High variability of speaking rates and styles
- Grammar rules of written language not strictly followed
- Hesitations: *ah, uhs, ums*
- Filler words and phrases: *yeah, you know*
- False starts, aborted or stuttered words: *let's meet mon- no tuesday*
- Non verbal sounds: *breathing, tongue clicks*
- Limited frequency bandwidth, noisy audio channels

iARPA Babel Project

www.language-service.co.nz

Apply lingustic, machine learning, and speech processing methods to enable speech recognition for keyword search (Harper, ASRU 2013, Coling 2014)
http://www.iarpa.gov/index.php/research-programs/babel

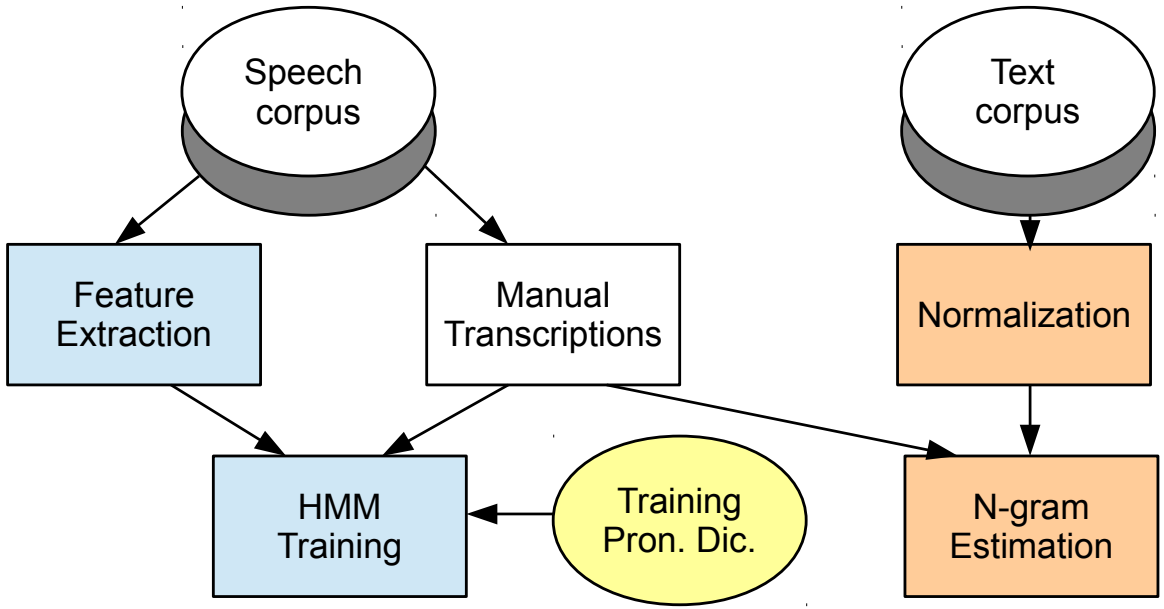
Lithuanian language

- 3.5 million speakers
- Baltic subgroup of Indo-European languages
- 2 dialects - Aukštaitian and Samogitian

Lithuanian language

- Based on Latin alphabet, 32 letters
- Some letters occur in recent loanwords
  1. *f* - filmas 'film', *ch* - chaosas 'chaos', *h* - humoras 'humor'
- Inflected language
  1. 5 of 11 parts of speech are inflective: noun, verb, adjective, numeral, and pronoun
  2. Words composed of root, stem, prefix, suffix, and ending
- Flexible word order in sentence
- 3 types of stresses, meaning of some words can be distinguished by stress

System training resources



- Acoustic model: telephone speech recordings with transcriptions
- Language model: written text
- Pronunciation dictionary
  1. Graphemes - easily derived
  2. Phonemes - linguistic skills, better represent speech production

Low resourced language

- Little acoustic data with corresponding transcriptions
- Text and untranscribed audio can be found on the Internet

Lithuanian phonemic inventory

Vowels	
a, ą	/a/, /ɑ/
e, ę, ė	/ɛ/, /æ/, /e:/
i, į, y	/i/, /i:/, /i:/
o	/o:/, /ɔ/
u, ų, ū	/ʊ/, /u:/, /u:/

- 32 letters
- 56 phonemes, 45 consonants and 11 vowels
- Consonants soft (palatalized) or hard (not palatalized)
- 8 diphthongs /ai/, /aʊ/, /ei/, /ʊi/, /oʊ/, /o:i/, /iɛ/, /ʊo:/
- 4 affricates /ts/, /ʧ/, /dz/, /dʒ/

Lithuanian phonemic inventory

	Consonants		
p	/p/,/p^{j}/	dz	/dz/,/dz^{j}/
b	/b/,/b^{j}/	č	/ʧ/,/ʧ^{j}/
t	/t/,/t^{j}/	dž	/dʒ/,/dʒ^{j}/
d	/d/,/d^{j}/	m	/m/,/m^{j}/
k	/k/,/k^{j}/	n	/n/,/n^{j}/
g	/g/,/g^{j}/	l	/ɫ/,/l^{j}/
v	/v/,/v^{j}/	r	/r/,/r^{j}/
s	/s/,/s^{j}/	j	/j/
z	/z/,/z^{j}/	f	/f/,/f^{j}/
š	/ʃ/,/ʃ^{j}/	ch	/x/,/x^{j}/
ž	/ʒ/,/ʒ^{j}/	h	/ɣ/,/ɣ^{j}/
c	/ts/,/ts^{j}/		

- Consonants are soft before the vowels /ɛ/, /æ/, /ė:/, /i/, /i:/
- Classes of sounds
  1. Voiced /b/, /d/, /g/, /z/, / ʒ/, /ɣ/, /dz/, /dʒ/, /v/, /j/, /m/, /n/, /ɫ/, /r/
  2. Voiceless /p/, /t/, /k/, /f/, /s/, /ʃ/, /x/, /ts/, /ʧ/

Grapheme to phoneme rules

Lithuanian has quite strong dependency between orthographic transcription and phonetic form

- *ia*, *iau*, *iai* pronounced *e*, *eu*, *ei*
  lauk**ia** /ɫaʊkɛ/ (waits)

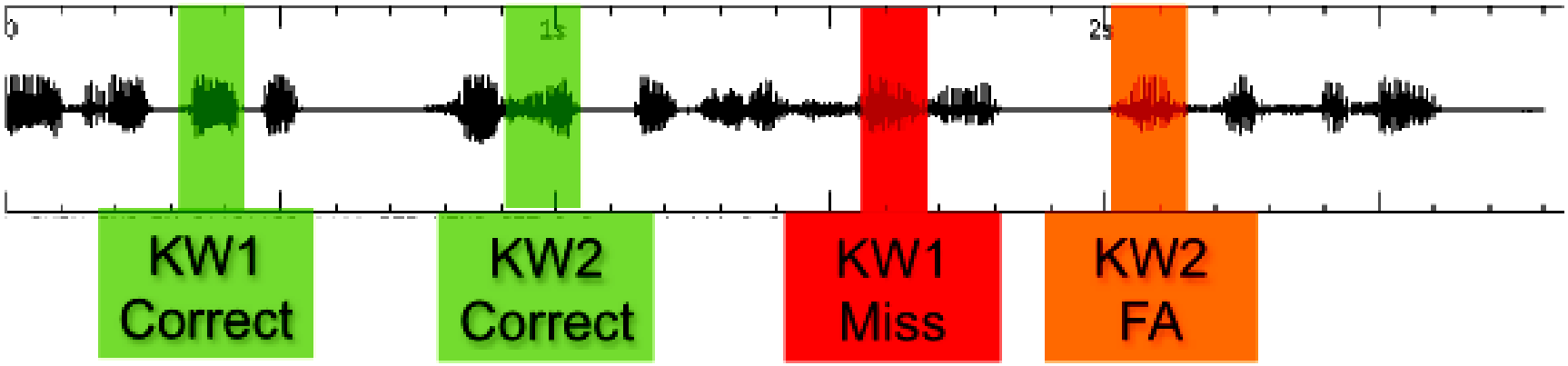
- voiceless consonants *p*, *t*, *k*, *s*, *š* before voiced *b*, *d*, *g*, *z*, *ž* are pronounced like voiced (anticipatory coarticulation)
  su**k**damas /sʊgdamas/ (turning)

- voiced consonants *b*, *d*, *g*, *z*, *ž* before voiceless *p*, *t*, *k*, *s*, *š* are pronounced like voiceless
  dir**b**ti /dirpti/ (to work)

- voiced consonants *b*, *d*, *g* at the end of word are pronounced like voiceless
  ka**d** /kat/ (that)

- ....

Keyword spotting system



- Keyword Spotting (KWS) - find only certain words
- Two types of errors - miss and false alarm
- Keywords
  1. Out-of-vocabulary (OOV) are missing keywords from the ASR system vocabulary
  2. In-vocabulary (INV)
- Performance metrics
  1. Actual term-weighted value (ATWV)
  2. Maximum term-weighted value (MTWV)

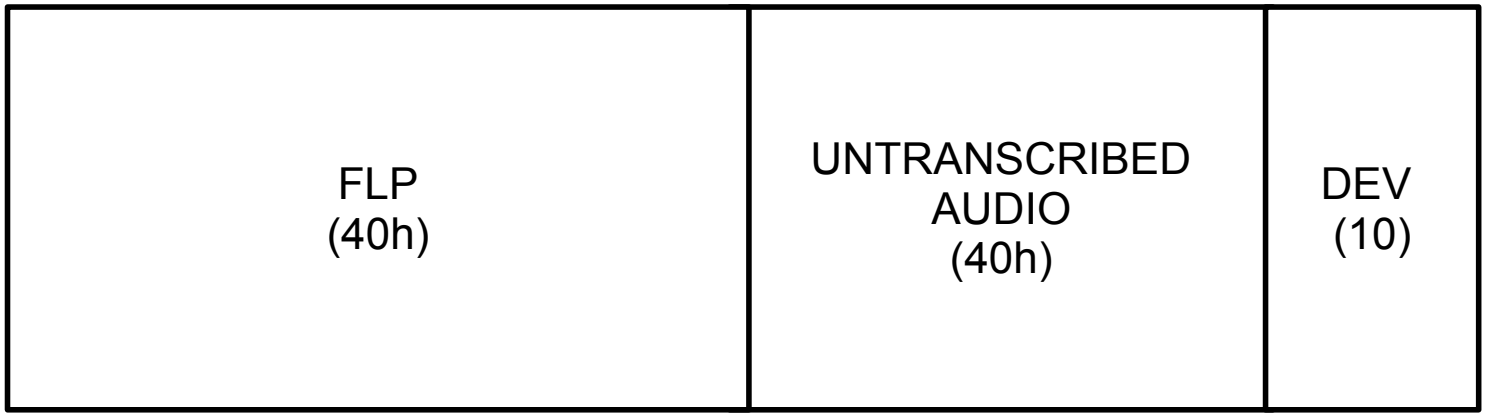
Data set

- 40 h with transcripts (+40 untranscribed) (Full Language Pack (FLP))
  1. IARPA-babel304b-v1.0b dataset
- 3 h trans (+77 untranscribed) (Very Limited Language Pack (VLLP))
- Web text corpora 26M word tokens
- Results reported
  1. Speech-to-text (STT) system on 10 hour dev set
  2. KWS on 4079 keywords for FLP condition, 10% OOV

Baseline recognition systems

- Speech-to-text system
  1. Left-to-right 3-state HMMs with Gaussian Mixtures
  2. Word position-dependent triphone-based models, tied-state
  3. Stacked bottleneck features, provided by Brno University of Technology [Grézl et al, 2013]
  4. Semi-supervised training (SST)
  5. 3-gram backoff LMs with Kneser-Ney smoothing
- Keyword spotting system
  1. Lattices converted to confusion networks
  2. Exact matches considered, though case and whitespace ignored
  3. Word and sub-word (character 7-gram) units [Hartmann et al, 2014]

FLP available data



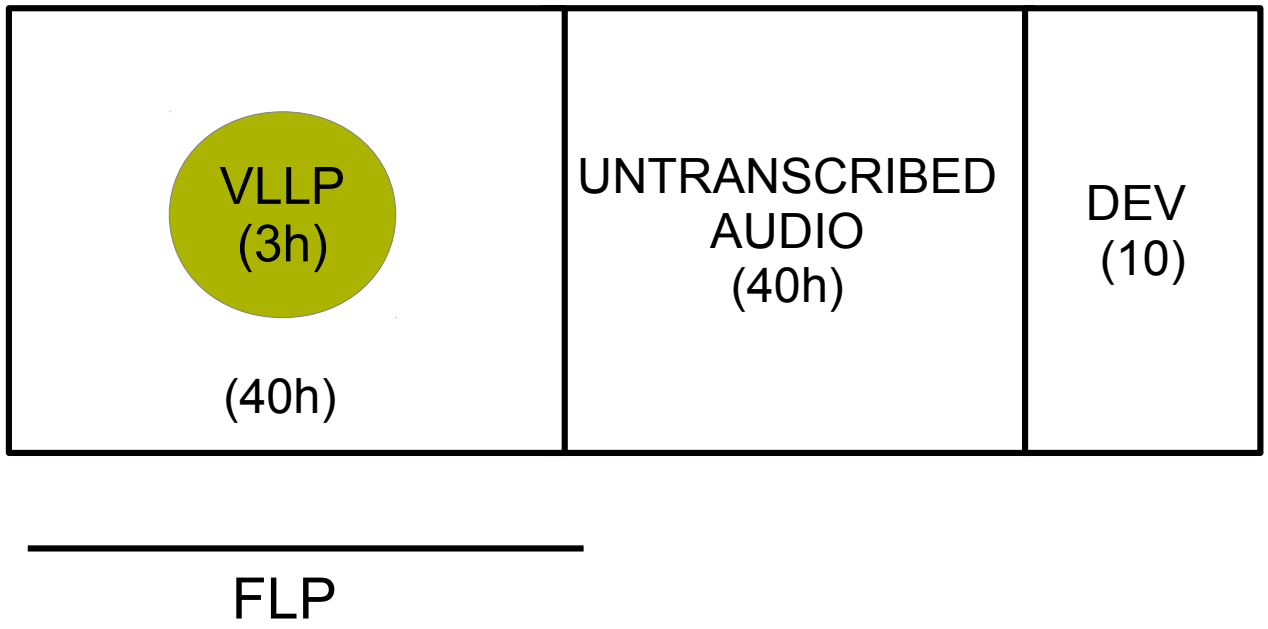
STT and KWS results for FLP systems

System	#Units	Modification from baseline
FLP graph-baseline	35	graphs
FLP phone-baseline	32	phones
FLP phone	36	affricates
FLP phone	38	diphthongs [except ou, oi]
FLP phone	48	soft consonants [except soft ch]

System	#Units	%WER	MTWV(all/inv/oov)
graph-baseline	35	44.6	0.579/0.592/0.472
phone-baseline	32	44.7	0.576/0.591/0.476
phone	36	44.6	**0.580/0.593/0.487**
phone	38	**44.4**	0.576/0.591/0.460
phone	48	44.6	0.573/0.587/0.472

- Affricates separate unit, slight improvement in MTWV (36 phones)
- Diphthongs separate units, 0.3% decrease in WER (38 vs 32 phones)
- Best phone-system 0.2% absolute lower WER than graph-system

VLLP available data



WER results for VLLP systems

System	#Units	Modification from baseline
VLLP graph-baseline	33	graphs, c→ts, f→v
VLLP graph	29	z→s, ch→ɣ, ę→ɛ, į,y→i:, ū,ų→u:
VLLP phone-baseline	31	phones, f→v
VLLP phone	29	z→s, ch→ɣ, ę→ɛ, į,y→i:, ū,ų→u:

System	#Units	%WER
graph-baseline	33	52.6
graph	29	**52.2**
phone-baseline	31	52.3
phone	29	**52.0**

- Results slightly better when number of units reduced
- Best phonemic system about 0.2% lower WER than best graphemic system

WEB and SST

- In our experiments
  1. FLP→ manual transcriptions for LM
  2. VLLP → manual and WEB texts for LM, untranscribed data for AM
- WEB texts and SST help to reduce gap between FLP and VLLP
- What is impact of WEB texts and SST for both FLP and VLLP systems?

WER results for different conditions

Set	Hours	AM	LM	Lexicon	%OOV	%WER
FLP	40	trn	trn	30k	7.6	44.4
FLP	73	trn + SST	trn	30k	7.6	44.8
FLP	40	trn	trn + web	60k	5.2	**42.4**
FLP	73	trn + SST	trn + web	60k	5.2	42.4
VLLP	3	trn	trn	5.7k	16.7	59.3
VLLP	41	trn + SST	trn	5.7k	16.7	59.0
VLLP	3	trn	trn + web	60k	6.0	53.3
VLLP	41	trn + SST	trn + web	60k	6.0	**52.0**

- FLP 40 vs VLLP 3: reduces absolute WER by 15%
- Web texts: for VLLP the WER is reduced by 6%, under 2% for FLP
- SST improves VLLP both with and without Web data
- Best VLLP WER remains 10% behind of FLP

Conclusions

- Developed conversational telephone speech recognition system for Lithuanian, a low-resourced language
- Compared phoneme-based and grapheme-based systems
  1. Phonemes give only a slight improvement for two training conditions (3 or 40 hours of transcribed audio data)
  2. Strong relationship between the orthographic and phonemic forms in Lithuanian
- Explored impact of Web texts for training language models, and untranscribed data for semi-supervised training of acoustic models
  1. Adding Web texts to FLP system gave an improvement of about 2%
  2. VLLP system was improved more 7% absolute using both Web data and semi-supervised training

Thank you for your attention!

Acknowledgments. *We would like to thank our IARPA-Babel partners for sharing resources (BUT for the bottle-neck features and BBN for the web data), and Grégory Gelly for providing the VADs. This research was in part supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.*