Advances in Estonian Spoken Language Technology

Tanel Alumäe

Laboratory of Phonetics and Speech Technology
Institute of Cybernetics
Tallinn University of Technology
Estonia

Final Workshop of CDC 2002-2007

Outline

Introduction

Spoken Language Technolgy

Subfields

- Automatic speech recognition (speech-to-text)
- Speech synthesis (text-to-speech)
- Spoken language understanding
- Automatic speech-to-speech translation

Interdisciplinary field

- Acoustics
- Phonology
- Phonetics
- Linguistics
- Semantics
- Psychology
- Computer science

Motivation

Spoken Language Technolgy

Applications

- Speech-based and multimodal interfaces
- Automatic dictation systems
- Automatic dialogue systems
- Spoken data retrieval systems
- Speech transcription and summarization systems
- Automatic speech translation systems

Motivation of research

- Language technology is essential for language survival
- Estonian is a very small language
- No commercial interest in Estonian language technology development from companies
- Subsidiarity principle does not allow the EU to provide financial support for HLT development of smaller languages

Spoken Language Technology in Estonia

Language technology actions in Estonia

- The status of the Estonian language and it's protection is validated by the constitution (since 2007)
- Development Strategy of the Estonian Language 2004-2010
- National Estonian Language Technology Programme 2006-2010

National Estonian Language Technology Programme 2006-2010

- **Main goals**: development of language resources and language-specific human language technology modules
- **Financing**:
  - ca. 7M EEK per year, i.e. ca €450K per year (2006 and 2007)
  - 17M EEK, i.e. ca €1.1M for 2008 (state budget plan)
- **Key players**:
  - University of Tartu
  - Institute of the Estonian Language
  - Institute of Cybernetics at TUT
  - Filosoft Ltd

National Estonian Language Technology Programme 2006-2010

On-going projects: (2006 – 17 projects, 2007 – 20 projects):

- **Speech corpora** – emotional speech, spontaneous speech, dialogues, non-native speech, etc.
- **Text corpora** – written language corpus, multi-lingual parallel corpora, etc.
- **Research/technology development** – speech recognition, speech synthesis, machine translation, information retrieval, lexicographic tools, syntactic analysis, semantic analysis, dialogue modeling, etc

Development of human resources

- **Doctoral School of Linguistics and Language Technology at the University of Tartu (2005-2008)**
  - Main goals:
    - ★ improve the quality of doctoral studies in linguistics and language technology
    - ★ prepare 20 new PhDs
  - Partners:
    - ★ Institute of the Estonian Language
    - ★ Institute of Cybernetics at TUT
    - ★ Several foreign universities and local industrial partners
- Curricula on Computer Linguistics at Tartu University
- Speech technology courses at Tallinn University
- International cooperation – e.g. NGSLT, NordForsk networks

Highlights of recent advances in spoken language technology

- Rich set of supporting tools and technologies
  - Morphological analysis and synthesis
  - Shallow syntax analysis
  - WordNet (provides semantic realtionships of words)
- Active work on collecting and transcribing various corpora
  - Corpus of news broadcasts from Estonian Radio
  - Dictated speech corpus for HQ speech synthesis
  - Corpus of emotianal speech
  - Corpus of dialogue act transcripts
  - Corpus of Estonian as a second language
  - Phonetic corpus of spontaneous speech
- Work on unit-selection based speech synthesis (much better quality than existing synthesis)
- Large vocabulary speech recognition is available as a prototype
- Propotype of the first spoken dialogue system (interface to a theatre information system)

Laboratory of Phonetics and Speech Technology

Laboratory of Phonetics and Speech Technology

Research fields

- Estonian phonetics
  - Estonian prosody and sound system
  - second language (L2) speech
- Speech technology
  - speech synthesis
  - speech analysis
  - speech and speaker recognition
  - phonetic databases

Current projects

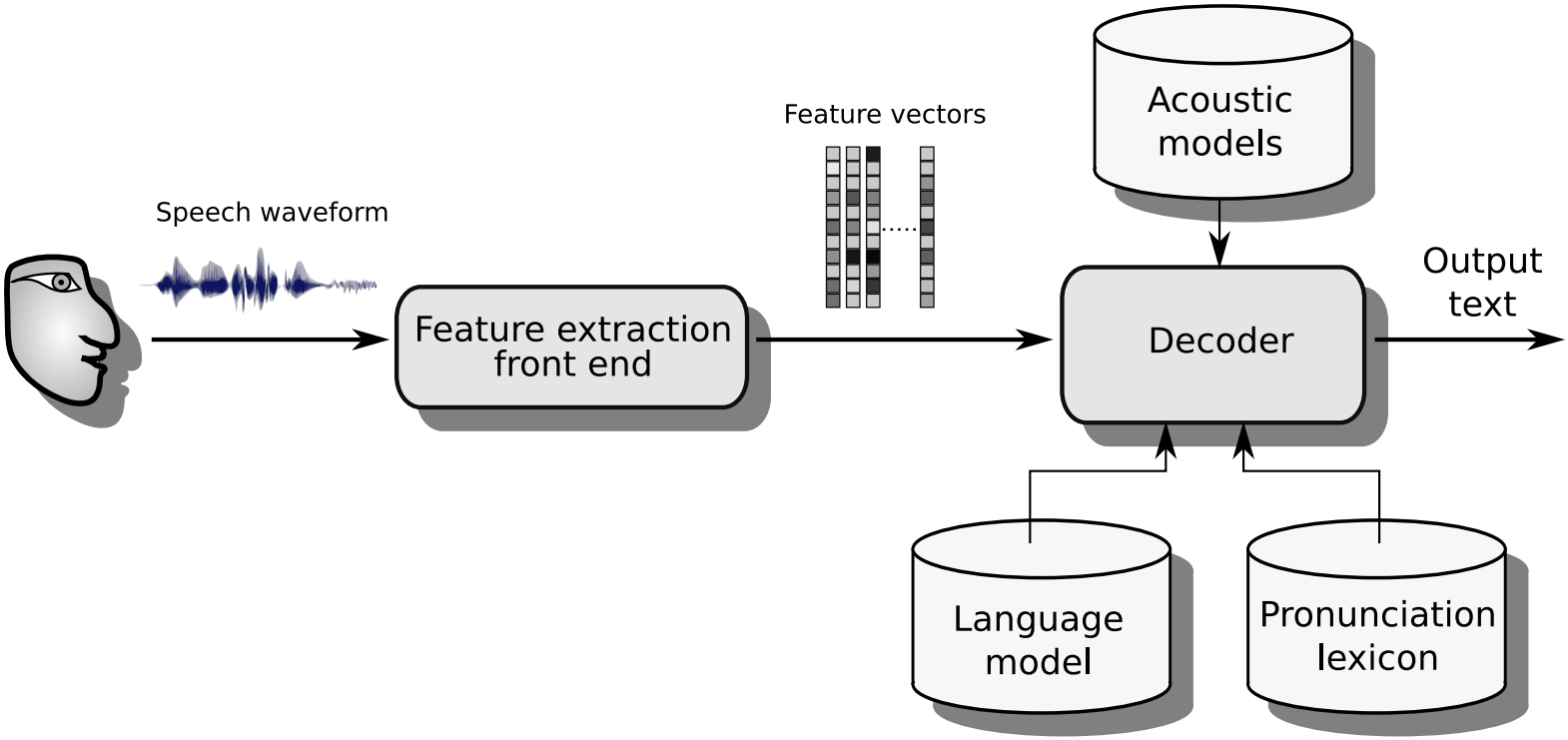
- Speech analysis and speech variability modelling
- Speech resources and databases
- **Research and development of methods for Estonian speech recognition**

Estonian speech recognition research

Estonian speech recognition research

Research goals

- Develop Estonian-specific methods and models for large vocabulary continuous speech recognition (LVCSR)
- Adapt modern statistical framework using hidden Markov models and *N*-gram language models, use available software



Statistical language modeling

- Used for calculating prior probabilities of words and sentences (world knowledge)
- Trained from very large text corpus
- The hardest problem for Estonian LVCSR
- Caused by the agglutinative, highly inflective and compounding nature of the language
  - Huge humber of different word forms
  - Word order relatively free
- Result: large out-of-vocabulary rate when using words as basic units for language modeling
- Solution: split words into morphemes using a morphological analyzer, use morpheme units for language modeling, e.g.

$$koolimajast \rightarrow kooli\ maja\ \_st$$

Language model adaptation

- LVCSR usually uses a general statistical language model trained on a mixed corpus
- However, speech is usually focused on a specific topic
  - e.g. news transcription: stories about inner politics, foreign issues, sports, weather
  - certain words co-occur often in certain topics
- **Language Model Adaptation**: given a few sentences as topic 'seed', adapt the general language model so that it predicts semantically related words with higher probability
- In LVCSR, morphemes are used as basic language units
  - Morphemes give high language **coverage**, given 60 000 most frequent units
- Are morphemes good units for LM adaptation?
  - Do morphemes carry enough semantic content?

Proposed approach

Outline of the proposed method:

- Use *latent semantic analysis* (LSA) for representing document 'closeness' measures
  - LSA uses large and very sparse word-document matrix co-occurrence matrix as input and applies truncated singular value decomposition (SVD) for dimensionality reduction. This extracts most characteristic components and ignores higher order effects (noise)
- Experiment with different language units (words, lemmas (base forms), or morphemes) for extracting semantic relationships
- Use short 'seed' text to find semantically close documents
- Use the morpheme unigram statistics in the closest documents to adapt the background morpheme LM

Experimental results

Speech recognition experiments

- Data: hourly short broadcast news from the national radio, manually segmented into stories and sentences
- For adaptation: $\sim$ 500 000 documents (mainly newspaper articles) were used for building topic models
- Experiment: run 1st pass using the general LM, use the recognized text for LM adaptation, and use the adapted LM in a 2nd recognition pass
- We measured letter error rate (LER) without and with adaptation
- Morpheme-based adaptation statistically significantly better

System	LER, %
No adaptation	7.1
Word-based adaptation	6.7 (-6%)
Lemma-based adaptation	6.6 (-7%)
Morpheme-based adaptation	**6.4 (-10%)**

Summary

Summary

- If we don't care of our languages no one will do it!
- Coordinated activities on national and on international level are important
- **Our lab:**
  - Active work on collecting speech corpora which are essential for modern speech technology development
  - Recognition of dictated speech available as a prototype
  - Statistical language modelling the most challenging area for Estonian LVCSR, work continues
  - Recognition of spontaneous speech is still far away (no corpora)