Creation of HMM-based Speech Model for Estonian Text-to-Speech Synthesis

Tõnis Nurk
Institute of the Estonian Language

05.10.2012 @ HLT 2012
Speech Synthesis

- Analogue for human reading
- Input – text; output – speech waveform
- Overview of a typical speech synthesis system:
Linguistic Processing

• Language specific
• Orthographic text converted into pronunciation text
• Linguistic context factors (phoneme, syllable, word, phrase, stress, accent, length etc)
• Vowel ‘e’ in ‘mees’ (‘*man*’ in Estonian)
  – preceding consonant ‘m’ (formant trajectories)
  – monosyllabic word (vowel duration and quantity)
Statistical Parametric Speech Synthesis (1/2)

• Based on hidden Markov models
• Speech described using parameters, rather than stored examples
• Parameters described using statistics (e.g., means and variances of probability density functions)
Statistical Parametric Speech Synthesis (2/2)

• HMM of a speech segment:
Overview of System HTS

1. SPEECH DATABASE
   - Speech signal
   - Excitation parameter extraction
     - Excitation parameter
     - Spectral parameter
   - Spectral parameter extraction
     - Spectral parameter
   - Training of HMM
     - Label

2. TEXT
   - Text analysis
     - Label
     - Excitation parameter
     - Spectral parameter
   - Parameter generation from HMM
     - Context dependent HMMs

3. SYNTHESIS FILTER
   - SYNTHESIZED SPEECH

4. SYNTHESIS GENERATION
Properties of Statistical Parametric Speech Synthesis

• Advantages
  – flexible (voice characteristics, speaking styles, emotions, speaker adaption)
  – robust against sparse data
  – small footprint, low computational resource need

• Drawbacks
  – low quality (vocoder, accuracy of acousting modelling, over-smooting)
Speech Corpus

- Necessary for training speech model
- Contextually labelled
- Phonetically rich and balanced
- Transcribed automatically
- Large amount of training data provides high-quality synthesized speech
- IEL’s Speech Corpus (ca 17 hours of speech from 5 speakers)
Linguistic Processing Unit

- Linguistic specification of the speech model must correspond to the capabilities of text analysis module.
- Text analysis modules developed under Festival
Creation of Speech Model

• Adapting HTS to Estonian
  – phonetic and phonological context factors
  (phoneme, syllable, word, phrase, stress, accent, length etc)
• Choosing training corpus
  – amount of data
  – phonetically balanced
• Test corpus
• Compatible with text analysis module
Evaluation of Speech Models

• Listening to synthesized test sentences
• Sentences of test corpus don’t contain in training corpus
• Different training corpora (from 100 to 2000 sentences)
• Different linguistic specifications (better results with smaller number of phonemes)
Quality of Synthesized Speech

- Intelligibility
- Pronunciation errors (mistakes by text analysis unit)
- Speech model quality is dependent on
  - high quality training corpus
  - text analysis unit
  - phonetic and phonological context factors
Examples

• Training corpus of 100 sentences – barely understandable
• Training corpus of 500 sentences – understandable
• „Harjumaa kolmeteistkümnnes tulemus geograafias on kõva tase.“
  – Liisi_lyh_250
  – Liisi_lyh_487
  – Liisi_500
  – Liisi_2000
  – Liisi_lyh_2000
  – Liisi_lyh_2000
  – Tõnu
Conclusion

• Statistical parametric speech synthesis is effective in synthesizing acceptable speech.
• Relatively small corpus to train a model on
• Speech models adaptable
• Future prospects