An Application Example for Unified Speech Synthesis and Recognition using Hidden Markov Models

Frank Duckhorn, Matthias Wolff, Guntram Strecha and Rüdiger Hoffmann

Birmingham, 30.03.2009
Structuring

- Project goal and implementation
- Speech recognition
- Speech synthesis
- Synthesis demonstration
Project goal

- we are developing a unified speech recognizer and synthesizer
- it will use identical, speaker independent Hidden-Markov-Models
- our final goal is to create a speech-based steering device for measuring devices in form of an USB-stick
- to achieve that we have to reduce resource requirement and power consumption of our algorithms

![Diagram showing headset, control device, and measuring device connected via Bluetooth and USB.](image-url)
Hardware setup

- we use a digital signal processor (DSP) with floating point support in combination with a field programmable gate array (FPGA)
- the FPGA will do mathematical costly algorithms and controls interfaces to headset (Bluetooth) and to the steered device (USB)
- the DSP will do algorithms with complex structures and has coordination functions
Working status

**Recognizer**
- up to now we have ported our speech recognizer as DSP program
- so we can do offline recognition test on DSP
- the FPGA implementation of some recognizing algorithms is at work

**Synthesizer**
- we have done the algorithmically concept for a synthesizer based on the same Hidden Markov Models like the recognizer
- more precisely it uses the mean vectors of Gaussians of speaker independent Hidden Markov Models
- the software implementation of that synthesizer works on PC systems, so we can demonstrate it
Speech recognition

**Recognition**
- **Speech Signal**
  - FFT + MEL-Filter
  - Delta + PCA
  - Density Calculation
  - Graph Search
  - Confidence
  - Recognition Result

**Recognition Parameter**
- 30 MEL Coefficients
- 24 Features
- 1024 Densities
- 1. + 2. Best Paths

**Training**
- Statistical Estimation
- HMM Training
- Composition + Minimize
- Recognition Network

**Training Parameter**
- 60h Speech Signals
- Mixture of 8 Gaussians per Emitting State
- 3-state HMM per Phoneme
- Lexicon as FST
- Grammar as FST

(identical for synthesizer)
Speech recognition

Speech recognition results

- free phoneme recognition (2:32 h speech signal): 59.5% correctness and 45.8% accuracy
- command recognition (20 commands, 1:35 h speech signal): 96.7% word recognition rate

Speech recognition processor time (DSP)

- grammar: 31 commands with different parameters (including up to 4-digit numbers)
- recognition network: 1654 states and 3476 transitions
- => processor time for graph search: 18.6 sec for 22 sec speech signal (85% real time factor)
  - (feature extraction and distance calculation is not prospected, because it will be done by the FPGA)
  - use algorithm: synchronous A* search
Speech synthesis

**Speech synthesis concept**

- speech synthesizer uses saved Gaussian index sequences for every morpheme
- that sequences are concatenated for the text to synthesize
- then the feature sequence is built by indexed Gaussian mean vectors
- to get rid of the speaker independent sound, we transform the features with a speaker adaptation matrix
- Gaussian index sequences are generated from spoken speech by labeling with the recognizer, where each morpheme must be spoken at least one time
- F0 and energy contours are saved for every morpheme like the Gaussian index sequence
Speech synthesis

**Synthesis**
- speech signal
  - LCQ-synthesis
    - 30 LCQ coef.
  - speaker adaption
    - 30 MEL coef.
  - inverse PCA
    - 24 features
  - mean vector selection
    - morpheme seq.
  - morpheme splitting
    - text

**Synthesis parameter**
- F0 and energy seq. per morpheme
  - speaker adaption matrix
    - Gaussian index seq. per morpheme
  - PCA-Matrix
  - Gaussians

**Training**
- analyze
  - statistical estimation
  - automatic labeling by recognizer
    - 8 per mix
  - 3-state HMM per phoneme
    - identical for recognizer

**Training parameter**
- one speech signal of every morpheme
- 60h speech signals

**Grammar**

---

An Application Example for Unified Speech Synthesis and Recognition using Hidden Markov Models

TU Dresden, 30.03.2009
Speech synthesis demonstration

Demonstration configuration

- Project goal is to create a German speaking steering device, so we synthesize German sentences.
- Grammar: 8 prompt types + different parameter + up to 4-digit numbers.
- 39 speakable morphemes (distinction in start, mid and end => 79 different units).
- Sequences (Gaussian + F0 + energy) for every unit from 3 different speakers.
- Speaker adaption matrix for 3 different speakers.
- => So we can mix prosody from one speaker with adaption from another.
Thank you for your attention!