Language resources for the adaptive speech synthesis of dialects

Telecommunications Research Center Vienna (FTW)&

Acoustics Research Institute (ARI)+

7. Kongress der int. Gesellschaft für Dialektologie und Geolinguistik

Tuesday 24th June 2012
State-of-the-art

- Intelligibility of synthetic speech (solved)
  - diphone based speech synthesis, formant synthesis
- Naturalness of synthetic speech (solved)
  - unit selection based speech synthesis
- Flexibility of TTS systems (solved)
  - HMM based speech synthesis
- Conversational speech synthesis (unsolved)
  - System that speaks like in a natural human-human conversation in any speakers voice (variety switching, prosody, non-linguistic particles (filled pauses, hesitations, laughing, whispering))
Applications

- Web reader (http://wien.at)
- Screen reader for blind users
- Spoken dialog systems
  - Call center automation
  - Information systems (Viennese dialect dialog system 01/8904055-7051)
- Multimodal dialog systems
  - Car navigation systems
  - Personal digital assistant (Siri)
  - Virtual reality applications
Persona design for speech-based interfaces

- “there is no such thing as a voice user interface with no personality” (Cohen, et.al. 2004).
- Perception of sociolect and dialect influence our evaluation of speaker’s attributes (competence, intelligence, friendliness, etc.).
- Persona is defined as the “Standardized mental image of a personality or character that users infer from the application’s voice and language choice” (Cohen, et.al. 2004).
- Speech synthesis is an essential part of a spoken dialog system’s persona.
A text-to-speech synthesis system consists of:

1. **Text analysis**: Numbers, abbreviations, etc.
2. **Grapheme-to-phoneme conversion**
   - dictionary look-up
   - decision tree based grapheme-to-phoneme rules
3. **Prosody prediction** (pauses, durations, F0) and waveform generation
   - Concatenative: Unit selection speech synthesis
   - Parametric: Hidden Markov model (HMM) based speech synthesis
   - Concatenative and parametric: Hybrid systems
Decision tree for G2P conversion of German

Decision trees for G2P conversion of Standard German

Decision tree for letter y

- n.n=t
- true
- p=s
- Y
- i:
- n.n=e
- y:
- Y
- n=n
- Y
- j

Decision tree for letter x

- k s

Decision tree for letter w

- n=\#
- true
- n=l
- \epsilon
- n=s
- \epsilon
- n=v
- \epsilon
- v
Speaker independent (adaptive) HMM based speech synthesis system

- Training of models for spectral (Mel-cepstrum), excitation parameters (F0), and duration.
- Adaptation of models with target speaker data.
- Generation of parameters from adapted models.
- Synthesis from parameters.
- Austrian German voice adapted with 200 utterances audio2/at-adapt1.wav, audio2/at-adapt2.wav (Voices in Edinburgh HTS library 0.99).

Figure: Adaptive HMM-based speech synthesis system.
Speaker independent (adaptive) HMM based audio-visual speech synthesis system

Figure: Adaptive HMM-based audio-visual speech synthesis system.
Speaker adaptive training (SAT)

- If we already know that a model will be used for adaptation we can apply adaptation specific training strategies like speaker adaptive training (SAT).
- The goal in SAT is to estimate a HMM $\lambda$ such that the transformations $W_1(\lambda), \ldots, W_8(\lambda)$ maximize the likelihood of the adaptation data $O_1, \ldots, O_8$ (8 different speakers).
Context clustering

- To deal with unseen data (i.e. unseen quinphones) decision-tree based clustering is performed where the whole possible feature space is clustered.

- Acoustic-articulatory features can be used for clustering.

- In shared decision-tree clustering we train one decision tree per state (mostly used in synthesis).

- In phonetic decision-tree clustering we train one decision tree per state and phone (mostly used in recognition).

Figure: Decision-tree based state tying.
Figure: Part of decision-tree for mel-cepstrum of 3rd state (central state in 5-state HMM) for variety independent / speaker dependent model with full feature set.
Building TTS systems from scratch

1. Defining the phone set of the language / variety / dialect.
2. Create a recording script.
3. Selection of appropriate speakers.
4. Record the audio-visual data.
5. Automatically align the data.
6. Build the utterance data structure including syllabic (stress) and prosodic information.
7. Train the voice models (unit selection, HMM-based, or hybrid).
8. Develop the front-end including text analysis, lexicon, and grapheme-to-phoneme rules.
9. Develop interfaces for integration.
Defining the phone set of the language / variety / dialect

- Defining a phone set for a new language needs special linguistic knowledge.
- In one of our previous projects we have defined phone sets for Viennese varieties (https://portal.ftw.at/projects/vsds).
- In this project we have defined phone sets for
  - the dialect of Bad Goisern, Upper Austria (South-Middle Bavarian transition zone) and
  - the dialect of Innervillgraten, Eastern Tyrol (South-Bavarian dialect) (https://portal.ftw.at/projects/avds).
- Direct correspondence of phones and particular parts in the speech signal - necessary for automatic alignment
Create a recording script

- Compilation of a set of 600-700 phonetically transcribed sentences
- The sentences have to be phonetically balanced with respect to
  - the phone set established for the dialect
  - frequency of occurrence of each phone in the data
  - sufficient context-specific variation of phones
- The sentences are extracted from a larger corpus of material
  - 18-20 hours of recordings for each dialect, at least 10 speakers / dialect
  - spontaneous speech (elicited with key words) and translation tasks
- Creating a lexicon of occurring words in the material
- The sentences are recorded with 4 speakers (2 male, 2 female) for each dialect
Selection of appropriate speakers

Linguistic criteria

- “Native speaker”
- Consistent application of characteristic phonological processes (e.g. assimilations, deletions)
- Lexical knowledge and morpho-syntactic competence

Non-linguistic criteria

- Readiness to participate
- Concentration capacity
- Physiological characteristics (beard, eyes, glasses)
Record the audio-visual data - Hardware for visual recordings

- 6 infrared (IR) cameras.
- 1 grayscale video camera.
- Synchronization hub.
- Markers and calibration equipment.

**Figure:** Hardware for visual marker recording.
Record the audio-visual data - Visual features for marker-based synthesis

Visual speech is characterized by the movements of 42 marker points in the face.

audio2/psc_ivg_019-050.avi

Figure: Audio-visual recordings of Innervillgraten speaker.
Record the audio-visual data - Read and play prompts

For recording the audio-visual dialect data we used a setting where

- the speaker can hear the utterance the he/she is supposed to say
- and at the same time see an orthographic transcription of the utterance.


Na vorgestern bin ich in einem Haus gewesen, na garstig und dreckig ist es da gewesen, na fürchterlich. (audio2/p271_006.wav)

This is not necessary

- when an orthographic standard is available
- and the speakers know how to produce speech from the standard transcription.


Gestern stürmte es noch. (audio2/mpu_BERLIN_005.wav)
Synthesis samples

- Acoustic Viennese speaker dependent voice (http://cordelia.ftw.at/index3.html)
- Acoustic East Tyrolean (Innervillgraten) speaker dependent voice
  - Recorded - audio2/lsc_ivg_497.wav, audio2/lsc_ivg_508.wav
  - Synthesized - audio2/lsc_ivg_497_synth.wav, audio2/lsc_ivg_508_synth.wav
- Audio-visual adapted Austrian German voices (http://userver.ftw.at/~schabus/interspeech2012/)
Conclusion

- We showed a state-of-the-art audio-visual speech synthesis system.
- We discussed the importance of realistic personas for spoken dialog systems.
- We showed how to perform speaker selection, phone set definition, and recording for synthesis of varieties.
- In future work we will investigate adaptive audio-visual modeling of 2 dialects and transformation of varieties.
References and Acknowledgements


- This work was funded by the Austrian Science Fund (FWF): P22890-N23.