Michael Pucher, Dietmar Schabus, Gregor Hofer

From Viennese to Austrian German and back again - An algorithm for the realization of a variety-slider

Telecommunications Research Center Vienna (FTW)

7. Kongress der Internationalen Gesellschaft für Dialektologie und Geolinguistik (SIDG)

Tuesday 24th June 2012
What do we mean by “variety”?

- We are mainly interested in social (sociolects) and regional varieties (dialects).
  - where sociolects are defined by education and age.
  - and dialects are defined by geographic region.

- But the interpolation methods can also be applied to other language variations like accents of non-native speakers.
Interpolation - Why it works

- Interpolation can be best explained in the low-dimensional space of F1 and F2.
- Linear interpolation of formants leads to non-linear changes in the auditory-articulatory space.
- Mel-frequency cepstral coefficients (MFCC) are a representation of the spectral envelope.
- We interpolate between MFCCs to interpolate the whole acoustic model.

**Figure:** Hans-Heinrich Wängler, *Atlas deutscher Sprachlaute*, Akademie-Verlag Berlin 1981.
Interpolation - Problems

- But interpolated vowels (in-between) vowels may not exist.
- In-between vowels may not exist in a specific linguistic context (e.g. [ə] in stressed position in German).
- The transition between consecutive phones may be different in speed.
- This can lead to linguistically invalid models.
- Restriction of interpolation to certain phonological processes.
Speaker dependent Hidden Markov model (HMM) based speech synthesis system

- Training of models for spectral (Mel-cepstrum), excitation parameters (F0), and duration.
- Generation of parameters from trained models.
- Synthesis from parameters.
- Austrian German voice trained with 3000 utterances
  audio1/at-spkdep.wav (Voices in Edinburgh HTS library 0.99).

**Figure:** Speaker dependent HMM-based speech synthesis system.
Interpolation - Methods

Interpolation of HMMs can be used for
- generation of emotional speech
- generation of in-between varieties
- generation of fast speech.

Interpolation methods (Yoshimura et al. 2000):
- Interpolation among observations

\[ \hat{\mu} = \sum_{k=1}^{n} a_k \mu_k, \quad \hat{\Sigma} = \sum_{k=1}^{n} a_k^2 \Sigma_k, \quad \sum a_k = 1, \quad (\hat{\mu} = a\mu_1 + (1 - a)\mu_2) \]

- Interpolation among output distributions

\[ \hat{\Sigma} = \sum_{k=1}^{n} a_k (\Sigma_k + \mu_k \mu_k^T) - \hat{\mu} \hat{\mu}^T \]

- Interpolation based on KL-divergence
Interpolation of varieties

- Interpolation between Austrian German and Viennese sociolect (Pucher et al. 2010).
- Alignment on phone level.
- One interpolation parameter for the whole utterance.
- Missing phones are mapped to source phone with zero duration.

Figure: Flow of dialect interpolation.
Schwa ([@], [ə])-deletion [gəviçt] - [gviçt]

audio1/und_mit_gewicht_bitte.wav
Schwa ([@], [ə])-deletion [gəvιkt] - [gviɛt]

audio1/und_mit_gewicht_bitte_inputswitch.wav
l-vocalization [milç] - [myːç]

---

<table>
<thead>
<tr>
<th>UND</th>
<th>MIT</th>
<th>MILCH</th>
<th>BITTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>1</td>
</tr>
<tr>
<td>0.8</td>
<td>L</td>
<td>1.2</td>
<td>1.4</td>
</tr>
</tbody>
</table>

audio1/und_mit_milch_bitte.wav
l-vocalization [milç] - [myːç]

audio1/und_mit_milch_bitte_inputswitch.wav
Input shift [ʃlaːk] - [ʃlɔːk]

<table>
<thead>
<tr>
<th>UND</th>
<th>MIT</th>
<th>SCHLAG</th>
<th>BITTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>a: 1</td>
<td>1.2</td>
<td>1.4</td>
<td>1.6</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>a: 1</td>
<td>1.2</td>
<td>1.4</td>
<td>1.6</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>a: 1</td>
<td>1.2</td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>A: 1</td>
<td>1.2</td>
<td>1.4</td>
<td>1.6</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>A: 1</td>
<td>1.2</td>
<td>1.4</td>
<td>1.6</td>
</tr>
</tbody>
</table>

audio1/und/mit/schlag/bitte.wav
Input shift [ʃlaːk] - [ʃlaːk]

<table>
<thead>
<tr>
<th>UND</th>
<th>MIT</th>
<th>SCHLAG</th>
<th>BITTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>a: 1</td>
<td>1.2</td>
<td>1.4</td>
<td>1.6</td>
</tr>
<tr>
<td>a: 1</td>
<td>1.2</td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>A: 1</td>
<td>1.2</td>
<td>1.4</td>
<td>1</td>
</tr>
</tbody>
</table>

audio1/und_mit_schlag_bitte_inputswitch.wav
Monophthongization [dɔetʃ] - [dæːtʃ]

<table>
<thead>
<tr>
<th>UND</th>
<th>MIT</th>
<th>DEUTSCH</th>
<th>BITTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8 OY</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8 OY</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8 OY</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8 ?</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8 3:]</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8 3:]</td>
</tr>
</tbody>
</table>

audio1/und_mit_deutsch_bitte.wav
Monophthongization [dɔɛtʃ] - [dætʃ]

audio1/und_mit_deutsch_bitte_inputswitch.wav
Evaluation results

- Listeners had to rate synthesized samples between 1 (‘strongly Viennese’) and 6 (‘strongly standard’).
- Interpolation without switching rules created a more continuous transition.
- Interpolation with switching rules led to a categorical difference.

Figure: Evaluation of similarity in terms of dialect. Multidimensional scaling is used for 2D visualization of evaluation results.
Interpolation of duration models

Interpolation between normal and fast speaking rate
(audio1/spo_rho_070_00.wav, audio1/spo_rho_070_14.wav, audio1/spo_spospo+f_070_14.wav) (Pucher et al. 2010).

Figure: Interpolation between normal and fast speech models.
Advanced interpolation

- The proposed method needs a manual alignment of phonemes.
- States are interpolated per phoneme.
- A more flexible approach finds an alignment between states automatically
  - using dynamic programming and similarities between acoustic models.
Applications

- **Spoken dialog systems**
  - Use language recognition to detect the speaker’s variety.
  - Select a voice in a variety that is closest to the speaker’s variety,
  - or select a variety that is related to the speaker’s variety.

- **Dialectology**
  - Comparison of spoken dialect maps and interpolated synthetic speech.
Conclusion

- We explained interpolation of varieties in formant space.
- Described a method for performing linear interpolation with HMM-based speech synthesis.
- Performed sample interpolations for several phonological processes of Austrian German / Viennese.
- Showed that interpolation is able to produce in-between varieties.


This work was funded by the Austrian Science Fund (FWF): P23821-N23, and by the Vienna Science and Technology Fund (WWTF).