Symbolic vs. acoustics-based style control for expressive unit selection

Ingmar Steiner\textsuperscript{1,2} Marc Schröder\textsuperscript{1} Marcela Charfuelan\textsuperscript{1} Annette Klepp\textsuperscript{1,2}

\textsuperscript{1}Language Technology Lab, DFKI GmbH

\textsuperscript{2}Department of Computational Linguistics & Phonetics, Saarland University

7\textsuperscript{th} ISCA Speech Synthesis Workshop
Kyōto, 2010
Background

Unit selection

+ high naturalness
- low flexibility

Expressivity as a “side-effect” of database design
Background

Unit selection

+ high naturalness
– low flexibility

Expressivity as a “side-effect” of database design

Motivation

Expressive unit selection with

• smooth joins
• correct style

from mixed-style database
The PAVOQUE expressive speech synthesis corpus

Prompt material

- 3,000 German sentences from WIKIPEDIA, optimized for coverage and prosodic variation
- 400 of these selected for optimal coverage for each style
- 150 style-specific extra prompts (per style)
The PAVOQUE expressive speech synthesis corpus

Recording and processing

- One male native speaker of German
- \( \sim 8.5 \) hours of speech (16 bit, 16 kHz)
- manually corrected phonetic segmentation
The PAVOQUE expressive speech synthesis corpus

Expressive styles

- neutral
  - “news-reading style”
- cheerful
  - “nice, optimistic, happy-go-lucky”
- depressed
  - “a wet blanket kind of person”
- aggressive
  - “aggressive, irritable and short-tempered”
- poker
  - “cool, laid back”
Overview

German unit selection voices built using DFKI’s open-source M\textsc{ary} \textsc{tts} text to speech platform (http://mary.dfki.de/)
Baseline voices

Built from PAVOQUE data, forced style control

4 519 utterances
Baseline voices

Built from PAVOQUE data, forced style control

- neutral (2,946 utts)

4,519 utterances
Baseline voices

Built from PAVOQUE data, forced style control

- neutral (2,946 utts)
- cheerful (393 utts)

4,519 utterances
Baseline voices

Built from PAVOQUE data, forced style control

- neutral (2,946 utts)
- cheerful (393 utts)
- depressed (393 utts)

4,519 utterances
Baseline voices

Built from PAVOQUE data, forced style control

- neutral (2,946 utts)
- cheerful (393 utts)
- depressed (393 utts)
- aggressive (394 utts)

4,519 utterances
Baseline voices

Built from PAVOQUE data, forced style control

- neutral (2,946 utts)
- cheerful (393 utts)
- depressed (393 utts)
- aggressive (394 utts)
- poker (393 utts)

4,519 utterances
Baseline voices

Built from PAVOQUE data, forced style control

- neutral (2,946 utts)
- cheerful (393 utts)
- depressed (393 utts)
- aggressive (394 utts)
- poker (393 utts)
- allstyles (4,519 utts)
Symbolic style target cost

allstyles with discrete target cost feature:

\[ x_{\text{style}} = \begin{cases} 0 & \text{if } \text{style}_{\text{target}} = \text{style}_{\text{cand}}. \\ 1 & \text{else} \end{cases} \]
Symbolic *style* target cost

allstyles with discrete target cost feature:

\[ x_{\text{style}} = \begin{cases} 
0 & \text{if } \text{style}_{\text{target}} = \text{style}_{\text{cand}}, \\
1 & \text{else}
\end{cases} \]

- symbolic voice (4,519 utts)
Acoustic style target cost

**allstyles** with continuous target cost feature based on voice quality parameter OQG\(^1\):

\[ x_{vq} = |vq_{target} - vq_{cand}| \]

\(vq_{target}\) predicted using CART

\(^1\)Lugger et al. (2006)
Acoustic *style* target cost

**allstyles** with continuous target cost feature based on voice quality parameter OQG\(^1\):

\[ x_{vq} = |vq_{target} - vq_{cand}| \]

\(vq_{target}\) predicted using CART

- **vq** voice (4 519 utts)

\(^1\)Lugger et al. (2006)
400 Wikipedia sentences resynthesized in each style:

- Smoothness baseline: allstyles voice
- Style match baseline: \( \langle style \rangle \) baseline voice \( \in \{\text{●, ○, ●, ○, ●, ○} \}\)
- Gold standard: original recordings
Dynamic utterance blacklisting

Candidates:

<table>
<thead>
<tr>
<th>f</th>
<th>n</th>
<th>o</th>
<th>r</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>$n_1$</td>
<td>$o_1$</td>
<td>$r_1$</td>
<td>$d_1$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>$n_3$</td>
<td>$o_3$</td>
<td>$r_4$</td>
<td>$d_4$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Resynthesize utt 1:
- candidates
Dynamic utterance blacklisting

Candidates

\[
\begin{align*}
\text{Candidates:} &\quad f_1 \rightarrow n_1 \rightarrow o_1 \rightarrow r_1 \rightarrow d_1 \\
&\quad f_2 \quad n_3 \quad o_3 \quad r_4 \quad d_4 \\
&\quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots 
\end{align*}
\]

Resynthesize utt 1:
- candidates
- utt 1 selected
Dynamic utterance blacklisting

Resynthesize utt 1:
- candidates
- utt 1 selected
- utt 1 blacklisted
Objective measures

Criteria

- **Style**: percentage of units selected from utterances with requested style

\[
RMSE_i = \sqrt{\frac{1}{P} \sum_{k=0}^{P-1} (g_i(k) - s_m(i)(k))^2}
\]
Objective measures

Criteria

- Style: percentage of units selected from utterances with requested style
- Smoothness: number of joins vs. number of units
Objective measures

Criteria

- Style: percentage of units selected from utterances with requested style
- Smoothness: number of joins vs. number of units
- Spectral distance from gold standard:

\[
RMSE_i = \sqrt{\frac{1}{P} \sum_{k=0}^{P-1} (g_i(k) - s_{m(i)}(k))^2}
\]
Effect of target cost feature weight (symbolic voice)
Effect of target cost feature weight (vq voice)

- **Other style**
- **Intended style**
- **Neutral**
Smoothness criterion

Mean span length (higher = fewer joins)

- forced-style
- all-styles
- symbolic
- vq
- vq+prosody
Smoothness criterion

Effect of \( \frac{\text{target cost}}{\text{join cost}} \) ratio on mean span length

(at target cost feature weight 100)
Spectral distance criterion

Effect of target cost feature weight on spectral distance to gold standard (aggressive style)
Discussion

• Unit selection voices built from mixed-style expressive database
• Two style target cost features:
  • symbolic (discrete)
  • acoustic (voice quality)
• Controlled variation of target cost weight and \( \frac{\text{target cost}}{\text{join cost}} \) ratio
• Symbolic control gives expected results
• Acoustic control complex (more features may improve results)
Outlook

Future work:

- Improve robustness of acoustic control with a mix of features
- Combine style selection with modification
- Perceptual evaluation