Neural Network Based Pitch Control for Various Sentence Types

Volker Jantzen
Speech Processing Group
TIK, ETH Zürich, Switzerland
Overview

- Introduction
- Preparation steps
  - Prosody corpus
  - Prosodic transcription
  - Phonetic segmentation of speech data
  - Extraction of pitch contour
- Neural network
  - Input / output coding
  - Architecture
  - Training algorithm
  - Training parameters
- Conclusions
  - Results
  - Future work
Introduction

- SVOX TTS System can only handle prosody of declarative sentences up to now
- Goal was to include prosody of
  - Different kinds of questions
  - Exclamations
  - Enumerations
  - Emphasis on words
- Prosody corpus was recorded in cooperation with LATL (University of Geneva) and Swisscom
- Pitch control in SVOX is done with a neural network
- A recurrent neural network was trained with data from the prosody corpus
Question Types

- Yes/No questions

  Braucht die Schweiz eine Kulturpolitik?

- Wh questions

  Doch was ist hier mit diesem “Wirken in den Dingen” gemeint?

- Alternative questions

  Hast du das Auto genommen oder bist du mit der Bahn gefahren?
Over 1600 German sentences spoken by the same female speaker who recorded the diphone corpus

The corpus consists of
- 858 Declarative sentences
- 585 Questions
  - 175 Wh questions
  - 335 Yes/No questions
  - 75 Alternative questions
- 227 Exclamations
- 71 Enumerations
The first 1000 sentences of the corpus were manually transcribed with regard to:

- **Accents**
  - 1 Main accent of the phrase
  - 2 Pitch accent
  - 3 Non pitch accent
  - 4 Secondary word accent
  - E Emphatic accent

- **Phrase boundaries**
  - / Short break
  - // Long break
  - /// Sentence boundary
Prosodic Transcription II

Phrase types

- P  Progredient phrase
- S  Semi-terminal phrase
- T  Terminal phrase
- Y  Question with rising pitch at the end
- W  Question with falling pitch at the end
- AI  Alternative question - initial phrase
- AM  Alternative question - middle phrase
- AF  Alternative question - final phrase
- LI  Enumeration - initial phrase
- LM  Enumeration - middle phrase
- LF  Enumeration - final phrase
- XM  Parenthetical phrase / extraposition on a medium pitch level
- XL  Parenthetical phrase / extraposition on a low pitch level
Prosodic Transcription III

Examples

<table>
<thead>
<tr>
<th>Word</th>
<th>Pitch</th>
<th>Duration</th>
<th>Silence</th>
</tr>
</thead>
<tbody>
<tr>
<td>hast</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>du:</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>Vir</td>
<td>2</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>klIC</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>g@</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>gla_upst</td>
<td>1</td>
<td>P /</td>
<td></td>
</tr>
<tr>
<td>di:</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>z@s</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>StYk</td>
<td>3</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>pa</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>pi:r</td>
<td>1</td>
<td>P /</td>
<td></td>
</tr>
<tr>
<td>za_i</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>?a_in</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>gYl</td>
<td>2</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>tI</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>g@r</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>fEr</td>
<td>0</td>
<td>. -</td>
<td></td>
</tr>
<tr>
<td>tra:k</td>
<td>E</td>
<td>Y</td>
<td>///</td>
</tr>
</tbody>
</table>

Neural Network Based Pitch Control for Various Sentence Types
**Phonetic Segmentation of Speech Data**

- Phonetic segmentation needed to find syllable nuclei within speech signal on which the pitch contour is computed.
- Forced alignment using Entropics HTK.
- Hidden Markov models that were trained on the phonetic corpus could be used on the prosody corpus without retraining.
- **Speech coding**
  - 16 kHz ESPS files
  - 25 ms Hamming windows
  - For each window: 12 MFCC, Energy, 12 $\Delta$MFCC, and $\Delta$Energy.
- **Architecture of HMMs**
  - One CHMM for each phone including glottal stop.
  - Left-to-right architecture.
  - 3 emitting states.
- No manual corrections of segmentation were necessary.
Extraction of Pitch Contour

- $F_0$ Computation done with ESPS procedure get_f0
- get_f0 uses autocorrelation
  - Frame step: 10 ms
  - Correlation window size: 7.5 ms
- Minimum $F_0$ set to 120 Hz, maximum $F_0$ set to 500 Hz
- Good results, virtually no octave jumps
Neural Network Based Pitch Control for Various Sentence Types

* Recurrent neural network with 2 hidden layers:

- Input layer: 56 + 10 nodes
- 1. hidden layer: 20 nodes
- 2. hidden layer: 10 nodes
- Output layer: 3 nodes
- 10 recurrent links from 2. hidden layer
**Input / Output Coding**

- For each syllable

  - **Input vectors**
    - 56 binary elements
    - Left context: $3 \times 3 = 9$ elements
    - Syllable in focus: 29 elements
    - Right context: $6 \times 3 = 18$ elements

  - **Output vectors**
    - 3 pitch values
    - Pitch at beginning, center and end of nucleus
    - Output range [0.2 .. 0.8] corresponding to [180 Hz .. 360 Hz]
Syllable- / Context Coding

Coding of syllable in focus
- Short / long vowel
- High / low intrinsic pitch
- Plosive before syllable nucleus
- Plosive after syllable nucleus
- (5) Accent type
- (10) Phrase type
- (3) Previous phrase boundary
- (3) Following phrase boundary
- Previous phrase progredient (ends with high pitch)
- Previous phrase semi-terminal (ends with medium pitch)
- Word boundary before syllable
- Word boundary after syllable

Coding of context
- Pitch accent
- Non pitch accent
- Break before / after syllable
Recurrent neural network with 2 hidden layers:

- Input layer: 56 + 10 nodes
- 1. hidden layer: 20 nodes
- 2. hidden layer: 10 nodes
- Output layer: 3 nodes
- 10 recurrent links from 2. hidden layer
Training algorithm

Output of each Neuron

\[ O_j = f\left( \sum_i W_{ji}O_i \right) \quad \text{with} \quad f = \frac{1}{1 + e^{-x}} \]

Training with backpropagation through time

- Backpropagation
  \[ \Delta W_{ji} = \eta \delta_i O_i \]

\[ \delta_i = \begin{cases} f'(\sum W_{ji}O_i) (D_i - O_i) & \text{for output neurons} \\ f'(\sum W_{ji}O_i) \sum_k W_{kj} \delta_k & \text{otherwise} \end{cases} \]

- Net is unfolded in time to regard additional error from recurrent links
Training

- Implementation in Matlab
- Utterances
  - Trainingset: 590 sentences
  - Testset: 200 sentences
  - Trainingset and testset have same distribution of sentence types
- Trainparameter
  - Learn rate: 0.1
  - Epochs: ca. 1000
- Control of training process
  - Predicted pitch contours were plotted against original pitch contours
  - Resulting pitch contours were imposed on original speech signals with PSOLA and listened to
Results

- Pitch Contour is linear interpolation of outputs
- Computed pitch contour is imposed on original speech signal with a PSOLA algorithm
- Natural durations and energy
- Examples
  - Declarative sentences
  - Exclamations
  - Yes/No questions
  - Wh questions
  - Alternative questions
Conclusions

- Typical pitch contours of the different question types were learned by the network
- Computed pitch contours are close to natural pitch contours
- Difficulties with sentences where main accent is on last syllable
- Enumerations have worse results than the other sentence types (fewest training data)
- Mean square error is not a good measure for naturalness
Future Work

- Further experiments to gather more experience about the behaviour of neural networks
- Find formal criteria to estimate the quality of a neural network
- Embed neural network into the SVOX System
- Adaptation of the syntax analysis of SVOX so that different question types can be analysed properly
- Use the prosody corpus to retrain the models used for duration control