Example-Based Automatic Phonetic Transcription
Language Resources and Evaluation Conference 2010

Christina Leitner, Martin Schickbichler, Stefan Petrik

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria

21 May 2010
Motivation

Why use automatic phonetic transcription?

- Phonetic transcriptions are an essential resource in speech technologies and linguistics.
  - Speech recognizers
  - Speech synthesis
  - Labelling of corpora
- Manual transcription is time-consuming, expensive and error-prone.
Motivation (2)

Benefits of automatic phonetic transcription

- Creation of draft transcriptions
  - Correction by human transcribers instead of creation from scratch
  - Faster and cheaper
- More objective than transcriptions of a team of human transcribers
- Consistency check of already transcribed material
Existing approaches

- Mostly based on Hidden Markov Models (HMMs)

- “Model-based”

![Diagram showing HMM parameters, Viterbi alignment, and optional language model]
Our approach

- Inspired by concatenative speech synthesis and template-based speech recognition

- "Example-based"

![Diagram showing the process of "Example-based" approach]

```
Database of examples

Candidate selection (opt.)

Pattern comparison

Synthesis

[akvaɾɛl]
```
Example-based APT

2 scenarios

- Constrained phone recognition

- Unconstrained phone recognition
Example-based APT

2 scenarios

- Constrained phone recognition
  - Decision based on audio sample and intermediate transcription derived from orthographic transcription by letter-to-sound rules

- Unconstrained phone recognition
Example-based APT

2 scenarios

- **Constrained phone recognition**
  - Decision based on audio sample and intermediate transcription derived from orthographic transcription by letter-to-sound rules

\[ \text{“Bäcker”} \rightarrow [\text{beke}] \]

- **Unconstrained phone recognition**
Example-based APT

2 scenarios

- **Constrained phone recognition**
  - Decision based on audio sample and intermediate transcription derived from orthographic transcription by letter-to-sound rules

  
  \[\text{“Bäcker”} \quad /b\ e\ k\ 6/ \quad \rightarrow \quad [b\ddot{e}k\ddot{e}]\]

- **Unconstrained phone recognition**
  - Decision based on audio sample only

  
  \[\rightarrow \quad [b\ddot{e}k\ddot{e}]\]
Example-based APT: system overview

Database of examples

- Three-phone speech samples
- Phone boundaries determined by doing forced alignment with the Hidden Markov Toolkit (HTK)
- 12 Mel Frequency Cepstral Coefficients (MFCCs) plus overall energy, delta and acceleration coefficients: 39 parameters per frame

Pattern matching

- Measure for similarity between two utterances
- Dynamic time warping (DTW) algorithm
- Segmental and open-begin-end DTW
Example-based APT: system overview (2)

Transcription synthesis

- Constrained phone recognition
  - Number of phones fixed
  - Most frequent phones from best matching three-phone samples

- Unconstrained phone recognition
  - Number of phones unknown
  - List of n best matching samples for each frame
  - Nearest neighbor classification
Example-based APT: system overview (2)

Transcription synthesis

- Constrained phone recognition
  - Number of phones fixed
  - Most frequent phones from best matching three-phone samples

- Unconstrained phone recognition
  - Number of phones unknown
  - List of n best matching samples for each frame
  - Nearest neighbor classification

“Bäcker” /b e k 6/

sil b e o k 6 sil
@ u
@ \ io
\ a
Example-based APT: system overview (2)

Transcription synthesis

- Constrained phone recognition
  - Number of phones fixed
  - Most frequent phones from best matching three-phone samples

- Unconstrained phone recognition
  - Number of phones unknown
  - List of n best matching samples for each frame
  - Nearest neighbor classification

“Bäcker” /b e k 6/

[bɛke]
Example-based APT: system overview (2)

Transcription synthesis

- Constrained phone recognition
  - Number of phones fixed
  - Most frequent phones from best matching three-phone samples

- Unconstrained phone recognition
  - Number of phones unknown
  - List of n best matching samples for each frame
  - Nearest neighbor classification

“Bäcker” /b e k 6/

\[ b \ e_o \ k \ 6 \]

[beke]

sil b b b e_o e_o e_o e_o k k 6 6 6 sil
Example-based APT: system overview (2)

Transcription synthesis

- Constrained phone recognition
  - Number of phones fixed
  - Most frequent phones from best matching three-phone samples

- Unconstrained phone recognition
  - Number of phones unknown
  - List of n best matching samples for each frame
  - Nearest neighbor classification

“Bäcker” /b ə k 6/

[beke]

sil b b b e_o e_o e_o e_o k k 6 6 6 sil

↓

b e_o k 6

[beke]
Evaluation

Evaluation database: ADABA

- Austrian pronunciation database
- 6 professional speakers: Austrian, German and Swiss
- Narrow transcriptions: 89 phonemes - instead of 45 in SAMPA German
- About 12,000 utterances per speaker (~ 5h speech)
- Recordings in studio quality

- Provided by Rudolf Muhr, Research Center for Austrian German
  http://adaba.at/
Evaluation (2)

Data set specification
- Restriction to a single speaker
- 85% training data, 5% development data, and 10% test data

Evaluation measures
- Percentage of correct phones and phone accuracy

\[ PC = \frac{N - D - S}{N} \times 100\% \]
\[ PA = \frac{N - D - S - I}{N} \times 100\% \]

- \( N \) ... total number of phones in the reference transcription
- \( D \) ... number of deletions, \( S \) ... number of substitutions
- \( I \) ... number of insertions.
Evaluation (3)

Benchmark: Comparison to a model-based transcriber

- Trained with Hidden Markov Toolkit (HTK)
- Same data and acoustic frontend
- 5-state left-to-right context-dependent triphone models with up to 16 GMMs
- For constrained phone recognition:
  Use of intermediate transcription for language model
Results

Constrained phone recognition

<table>
<thead>
<tr>
<th></th>
<th>Int. Tr.</th>
<th>Model-based</th>
<th>Example-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>83.36%</td>
<td>90.88%</td>
<td>91.95%</td>
</tr>
<tr>
<td>PA</td>
<td>81.22%</td>
<td>88.83%</td>
<td>89.89%</td>
</tr>
</tbody>
</table>

Performance differences are significant at the 0.1% level using the Matched-Pairs test.
Results

Constrained phone recognition

<table>
<thead>
<tr>
<th></th>
<th>Int. Tr.</th>
<th>Model-based</th>
<th>Example-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>83.36%</td>
<td>90.88%</td>
<td>91.95%</td>
</tr>
<tr>
<td>PA</td>
<td>81.22%</td>
<td>88.83%</td>
<td>89.89%</td>
</tr>
</tbody>
</table>

Performance differences are significant at the 0.1% level using the Matched-Pairs test.

Unconstrained phone recognition

<table>
<thead>
<tr>
<th></th>
<th>Model-based</th>
<th>Example-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>88.10%</td>
<td>85.21%</td>
</tr>
<tr>
<td>PA</td>
<td>86.96%</td>
<td>82.38%</td>
</tr>
</tbody>
</table>

Performance differences are significant at the 0.1% level using McNemar’s test.
Implementations

EXTRA

- Standalone Java application
  - Evaluation and analysis of transcriptions
  - Batch transcription mode

ELAN-EXTRA

- Extension for the ELAN linguistic annotation software

http://www.spsc.tugraz.at/people/stefan-petrik/project-extra
ELAN-EXTRA
Conclusion

- Example-based approach to automatic phonetic transcription
  - Comparison to concrete audio samples instead of model
  - Detection of rare pronunciation variants possible

- Useful support for transcription of speech corpora
  - Manual transcription of part of corpus - rest automatically
  - Consistency check easily feasible

- Evaluation on the ADABA database
  - Comparable to an HMM-based transcription system
  - Best results with a combination of rule-based and example-based APT
Discussion

Thank you for your attention!
References I


References II


Synthesis - constrained phone recognition
Synthesis - unconstrained phone recognition

Examples

- [sil a k]
- [b e_o k]
- [e_o k 6]
- [k a R]
- [e k 6]
- [6 R a]

Frames of input utterance

Frame Boundaries