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.

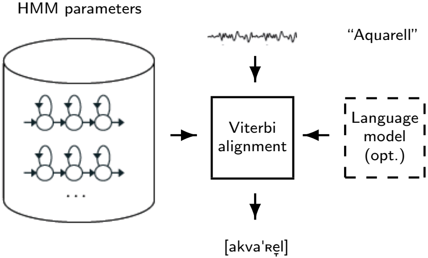
Motivaton (2)

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

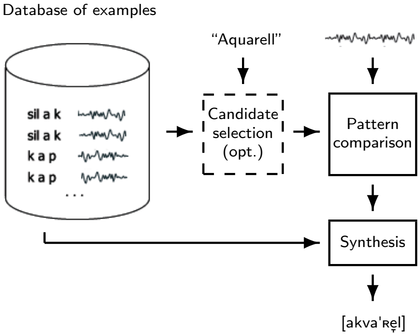
- Mostly based on Hidden Markov Models (HMMs)
  - "Model-based"



Our approach

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

- “Example-based”



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

 + "Bäcker" /b e k 6/ → [be̞kɐ]

- 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

 + “Bäcker” /b e k 6/ → [be̞kɐ]

- Unconstrained phone recognition
  - Decision based on audio sample only

 → [be̞kɐ]

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)

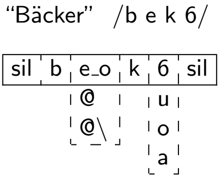
Transcription synthesis

- 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

Example-based APT: system overview (2)

Transcription synthesis

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

"Bäcker" /b e k 6/

b	e_o	k	6

[be̞kɐ]

- 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

"Bäcker" /b e k 6/

b	e_o	k	6

[be̞kɐ]

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

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

“Bäcker” /b e k 6/

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

| b e_o k 6 |

[be̞kɐ]

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

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

↓

| b e_o k 6 |

[be̞kɐ]

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 ($\sim$ 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

$$\text{PC} = \frac{N - D - S}{N} \times 100\% \qquad \text{PA} = \frac{N - D - S - I}{N} \times 100\%$$

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

	Int. Tr.	**Model-based**	**Example-based**
PC	83.36%	90.88%	91.95%
PA	81.22%	88.83%	89.89%

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

Results

Constrained phone recognition

	Int. Tr.	Model-based	Example-based
PC	83.36%	90.88%	91.95%
PA	81.22%	88.83%	89.89%

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

Unconstrained phone recognition

	Model-based	Example-based
PC	88.10%	85.21%
PA	86.96%	82.38%

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

Implementations

EXTRA

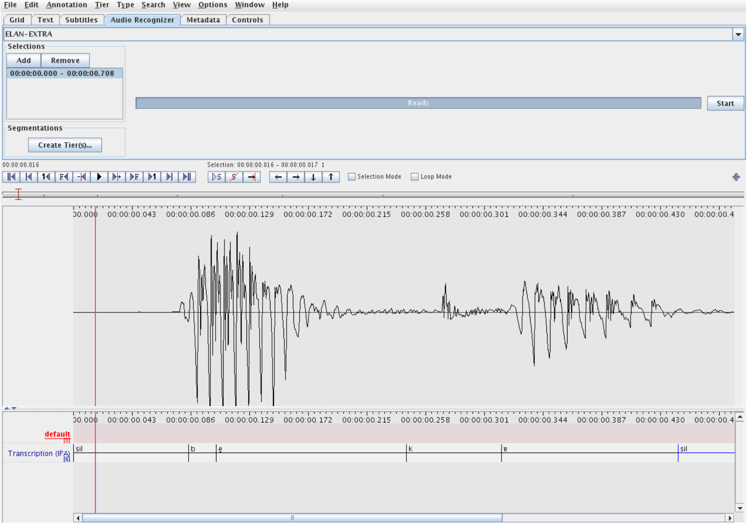
- Standalone Java application
  - Evaluation and analysis of transcriptions
  - Batch transcription mode

ELAN-EXTRA

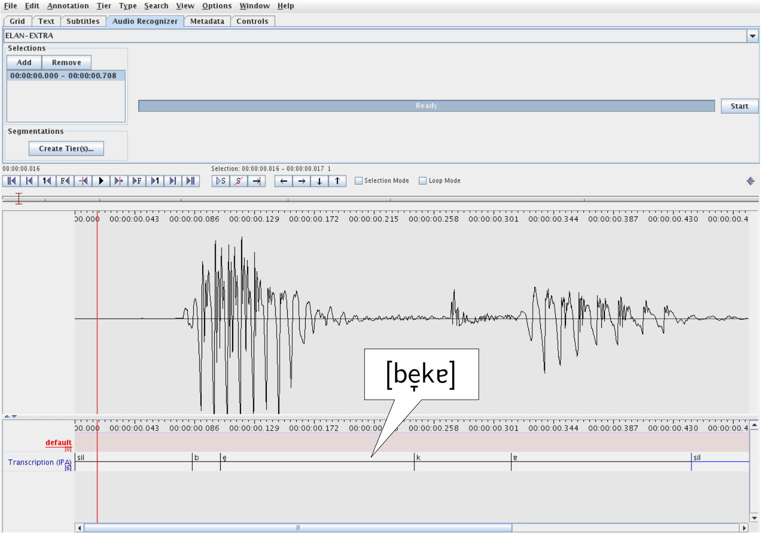
- Extension for the ELAN linguistic annotation software

`http://www.spsc.tugraz.at/people/stefan-petrik/project-extra`

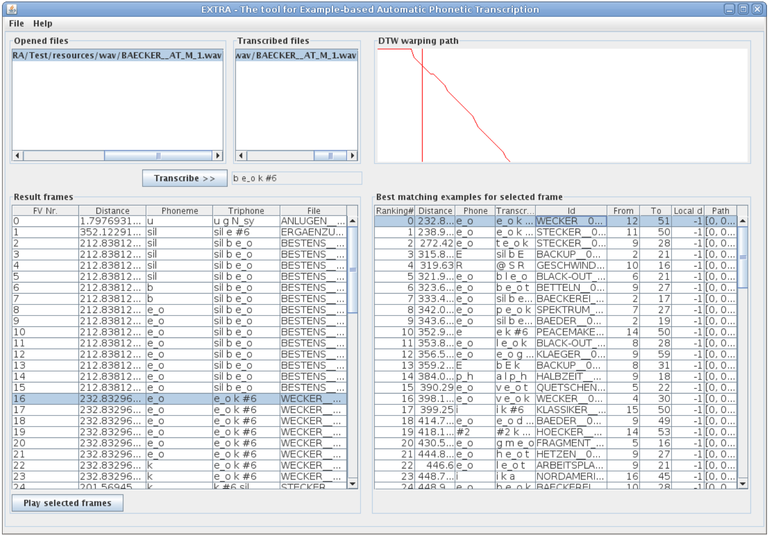
ELAN-EXTRA


File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Audio Recognizer Metadata Controls
ELAN-EXTRA
Selections
Add Remove
00:00:00.000 - 00:00:00.708
Ready
Start
Segmentations
Create Tier(s)...
00:00:00.016
Selection: 00:00:00.016 - 00:00:00.017 1
Selection Mode
Loop Mode
default
Transcription (IPA)
sil
b
k
sil

ELAN-EXTRA


File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Audio Recognizer Metadata Controls
ELAN-EXTRA
Selections
Add Remove
00:00:00.000 - 00:00:00.708
Ready
Start
Segmentations
Create Tier(s)...
00:00:00.016
Selection: 00:00:00.016 - 00:00:00.017 1
Selection Mode
Loop Mode
[be̞kɐ]
default
Transcription (IPA)
sil
b
e̞
k
ɐ
sil

EXTRA


EXTRA - The tool for Example-based Automatic Phonetic Transcription
File Help
Opened files
RA/Test/resources/wav/BAECKER_AT_M_1.wav
Transcribed files
wav/BAECKER_AT_M_1.wav
DTW warping path
Transcribe >>
b e_o k #6
Result frames
FV Nr. | Distance | Phoneme | Triphone | File
0 | 1.7976931... | u | u g N_sy | ANLUGEN_...
1 | 352.12291... | sil | sil e #6 | ERGAENZU...
2 | 212.83812... | sil | sil b e_o | BESTENS_...
3 | 212.83812... | sil | sil b e_o | BESTENS_...
4 | 212.83812... | sil | sil b e_o | BESTENS_...
5 | 212.83812... | sil | sil b e_o | BESTENS_...
6 | 212.83812... | b | sil b e_o | BESTENS_...
7 | 212.83812... | b | sil b e_o | BESTENS_...
8 | 212.83812... | e_o | sil b e_o | BESTENS_...
9 | 212.83812... | e_o | sil b e_o | BESTENS_...
10 | 212.83812... | e_o | sil b e_o | BESTENS_...
11 | 212.83812... | e_o | sil b e_o | BESTENS_...
12 | 212.83812... | e_o | sil b e_o | BESTENS_...
13 | 212.83812... | e_o | sil b e_o | BESTENS_...
14 | 212.83812... | e_o | sil b e_o | BESTENS_...
15 | 212.83812... | e_o | sil b e_o | BESTENS_...
16 | 232.83296... | e_o | e_o k #6 | WECKER_...
17 | 232.83296... | e_o | e_o k #6 | WECKER_...
18 | 232.83296... | e_o | e_o k #6 | WECKER_...
19 | 232.83296... | e_o | e_o k #6 | WECKER_...
20 | 232.83296... | e_o | e_o k #6 | WECKER_...
21 | 232.83296... | e_o | e_o k #6 | WECKER_...
22 | 232.83296... | k | e_o k #6 | WECKER_...
23 | 232.83296... | k | e_o k #6 | WECKER_...
Play selected frames
Best matching examples for selected frame
Ranking# | Distance | Phone | Transcr... | Id | From | To | Local d | Path
0 | 232.8... | e_o | e_o k ... | WECKER_0... | 12 | 51 | -1 | [0, 0...
1 | 238.9... | e_o | e_o k ... | STECKER_0... | 11 | 50 | -1 | [0, 0...
2 | 272.42 | e_o | t e_o k | STECKER_0... | 9 | 28 | -1 | [0, 0...
3 | 315.8... | E | sil b E | BACKUP_0... | 2 | 21 | -1 | [0, 0...
4 | 319.63 | R | @ S R | GESCHWIND... | 10 | 16 | -1 | [0, 0...
5 | 321.9... | e_o | b l e_o | BLACK-OUT_... | 6 | 21 | -1 | [0, 0...
6 | 323.6... | e_o | b e_o t | BETTELN_0... | 9 | 27 | -1 | [0, 0...
7 | 333.4... | e_o | sil b e... | BAECKEREI_... | 2 | 17 | -1 | [0, 0...
8 | 342.0... | e_o | p e_o k | SPEKTRUM_... | 7 | 27 | -1 | [0, 0...
9 | 343.6... | e_o | sil b e... | BAEDER_0... | 2 | 19 | -1 | [0, 0...
10 | 352.9... | e | e k #6 | PEACEMAKE... | 14 | 50 | -1 | [0, 0...
11 | 353.8... | e_o | l e_o k | BLACK-OUT_... | 8 | 28 | -1 | [0, 0...
12 | 356.5... | e_o | e_o g ... | KLAEGER_0... | 9 | 59 | -1 | [0, 0...
13 | 359.2... | E | b E k | BACKUP_0... | 8 | 31 | -1 | [0, 0...
14 | 384.0... | p_h | a l p_h | HALBZEIT_... | 9 | 18 | -1 | [0, 0...
15 | 390.29 | e_o | v e_o t | QUETSCHEN... | 5 | 22 | -1 | [0, 0...
16 | 398.1... | e_o | v e_o k | WECKER_0... | 4 | 30 | -1 | [0, 0...
17 | 399.25 | i | i k #6 | KLASSIKER_... | 15 | 50 | -1 | [0, 0...
18 | 414.7... | e_o | e_o d ... | BAEDER_0... | 9 | 49 | -1 | [0, 0...
19 | 418.1... | #2 | #2 k ... | HOECKER_... | 14 | 53 | -1 | [0, 0...
20 | 430.5... | e_o | g m e_o | FRAGMENT_... | 5 | 16 | -1 | [0, 0...
21 | 444.8... | e_o | h e_o t | HETZEN_0... | 9 | 27 | -1 | [0, 0...
22 | 446.6 | e_o | l e_o t | ARBEITSPLA... | 9 | 21 | -1 | [0, 0...
23 | 448.7... | i | i k a | NORDAMERI... | 16 | 45 | -1 | [0, 0...

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!

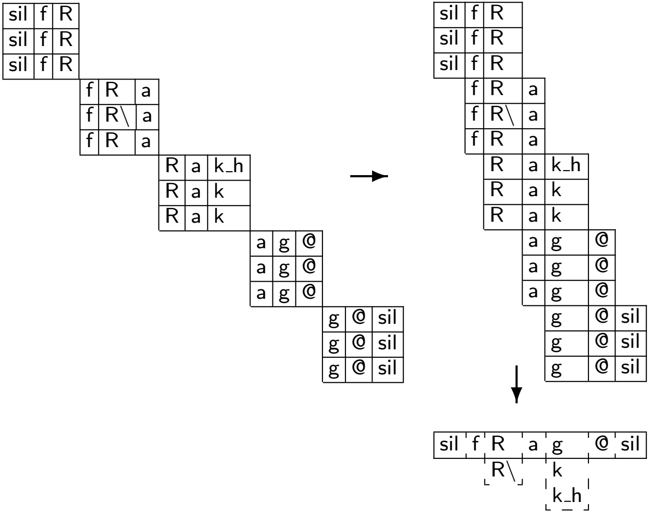
References I

- C. Cucchiarini and H. Strik, "Automatic phonetic transcription: An overview," *Proceedings of ICPhS*, pp. 347–350, 2003.
- M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1377–1390, 2007.
- C. Leitner, "Data-based automatic phonetic transcription," Master's thesis, Graz University of Technology, 2008.
- R. Muhr, "The Pronouncing Dictionary of Austrian German (AGPD) and the Austrian Phonetic Database (ADABA) – Report on a large phonetic resources database of the three major varieties of German," *Proceedings of LREC*, 2008.
- L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.
- A. Park and J. R. Glass, "Towards unsupervised pattern discovery in speech," *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*, pp. 53–58, 2005.

References II

- P. Tormene, T. Giorgino, S. Quaglini, and M. Stefanelli, "Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation," *Artificial Intelligence in Medicine*, vol. 45, no. 1, pp. 11–34, January 2009.
- P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a professional framework for multimodality research," *In Proceedings of Language Resources and Evaluation Conference (LREC)*, 2006.
- S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Engineering Department, 2006.

Synthesis - constrained phone recognition



Synthesis - unconstrained phone recognition