



Aalto University
School of Science

Stacked Transformations for recognizing foreign accented speech

Peter Smit

`peter.smit@aalto.fi`

`users.ics.tkk.fi/peter`

Speech Group

Adaptive Informatics Research Centre

May 18, 2010

About this presentation

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- Master's thesis presentation
- Supervisor: Docent Mikko Kurimo, D.Sc.(Tech.)
- Instructor: Janne Pyllkkönen, M.Sc.
- Work done in the Speech Group of the Adaptive Informatics Research Centre
- Funded by the EMIME project



Outline

Speech Recognition

Speech Recognition

Speaker Adaptation

Speaker Adaptation

Accented Speech

Accented Speech

Accent and
Neighbour
Transformations

Accent and Neighbour Transformations

Stacked
Transformations

Stacked Transformations

Experiments

Experiments

Conclusions

Conclusions



Speech Recognition

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- Good Speech Recognizers exist
 - Find contact in mobile phone
 - Telephone service that determines what customer representative you should be connected to
- Large Vocabulary Continuous Speech Recognition is harder
 - The larger the language model, the harder it is to recognize correctly.
 - On top of that, mismatch between training and evaluation



Speech Recognition - problems

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

'Mismatches' between recognizer and reality:

- Voices
- Noise Conditions
- Recording Equipment
- Speaking Style
- Vocabulary
- **Accents**



How does Speech Recognition work

Speech Recognition

Speaker Adaptation

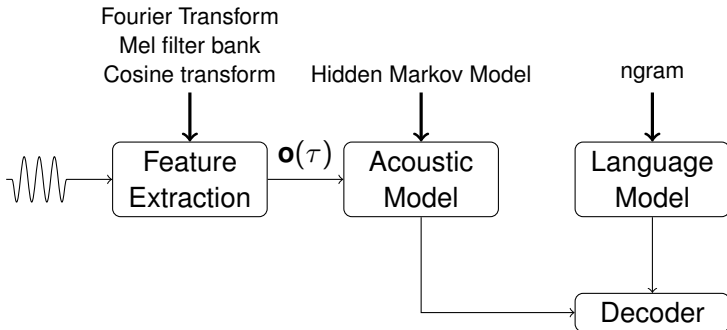
Accented Speech

Accent and Neighbour Transformations

Stacked Transformations

Experiments

Conclusions



Acoustic Model - Hidden Markov Model

Speech Recognition

Speaker Adaptation

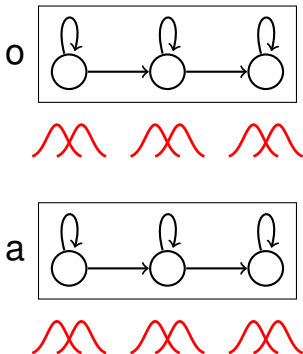
Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions



Speaker Dependent vs Speaker Independent

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- An ASR system is Speaker Dependent or Speaker Independent
- A SD system is more expensive, gives good performance for known speaker
- A SI system is less expensive, gives average performance for all speakers
- Adaptation transforms a SI model to a SD model with some adaptation data



Constrained Maximum Likelihood Linear Regression

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- Find a linear transformation for the means and / or variance of the emission distributions that maximizes the likelihood of some adaptation speech + transcription
- Use the same transform for the mean and the variance.
- Mean transform: $\hat{\mu} = \mathbf{A}^{-1} \mu - \mathbf{b}^{-1}$
- Variance transform: $\hat{\Sigma} = \mathbf{A}^{-1} \Sigma \mathbf{A}^{-1T}$



Regression Class Adaptation

Speech Recognition

Speaker Adaptation

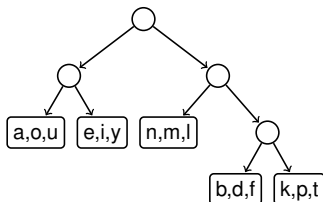
Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions



- Normally, all Gaussians are transformed with the same adaptation
- With Regression Classes, make an adaptation for a group of Gaussians
- More adaptation data means more Regression Classes

Recognizing Accented Speech

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- Best recognizer would be trained with accented speech
 - Accent Dependent model
 - Often not enough data available
- Adaptation
 - Dictionary
 - Acoustic Model adaptation
 - Cross-Lingual techniques



Accent Transformation

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- Accent data is used instead of speaker data
- More accent data is available, so more Regression Classes for adaptation
- Can be estimated before recognition



Neighbour Transformation

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- Find similar speakers (neighbours)
 - For example with eigenvoice parameters
- Use neighbours data for adaptation
- More data is available, so more Regression Classes for adaptation
- Most computations can be done before recognition



Stacked Transformations

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- Use first an Accent or Neighbour Transformation
- Still apply normal Speaker Adaptation
- Transformations must have different Regression Classes, else first linear transformations will be 'canceled' by the second.



Advantages Stacked Transformations

Speech Recognition

Speaker Adaptation

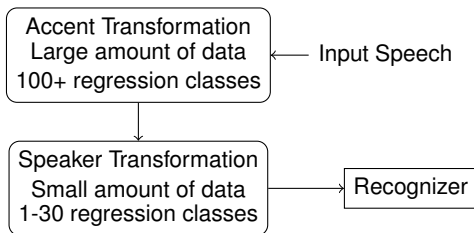
Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions



- No cost increase for recognition. First transform is calculated off line
- First transformation will give a better fit, therefore less speaker adaptation sentences will be needed for the same improvement

Speech Corpora

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

WSJ-84	American English speech
WSJCAM	British English speech
UED_F	Finnish accented English speech
UED_G	German accented English speech
UED_M	Mandarin accented English speech



Accented Speech in training

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

	No adap	Speaker adap
WSJ-84	40.6	29.4
WSJ-84+UED_F	33.9	25.2
WSJCAM	32.1	24.1
WSJCAM+UED_F	26.5	21.5

Table: Recognizing UED_F data, Word Error Rate %



Stacked Transformations

		WSJ-84	WSJCAM
Speech Recognition			
Speaker Adaptation			
Accented Speech	Baseline	40.6	32.1
	Speaker Adap (5 utt)	31.2	25.7
Accent and Neighbour Transformations	Speaker Adap (30 utt)	29.4	24.1
	5-Neighbour Adap	32.0	26.7
Stacked Transformations	5-Neighbour + Speaker Adap (5 utt)	28.5	24.5
Experiments	5-Neighbour + Speaker Adap (30 utt)	28.1	23.8
Conclusions	Accent Adap	29.0	26.4
	Accent Adap + Speaker Adap (5 utt)	26.7	23.8
	Accent Adap + Speaker Adap (30 utt)	26.4	23.8

Table: Finnish Accented Speech, Word Error Rate %



Stacked Transformations

		WSJ-84	WSJCAM
Speech Recognition			
Speaker Adaptation			
Accented Speech	Baseline	40.6	32.1
	Speaker Adap (5 utt)	31.2	25.7
Accent and Neighbour Transformations	Speaker Adap (30 utt)	29.4	24.1
	5-Neighbour Adap	32.0	26.7
Stacked Transformations	5-Neighbour + Speaker Adap (5 utt)	28.5	24.5
	5-Neighbour + Speaker Adap (30 utt)	28.1	23.8
Experiments			
Conclusions	Accent Adap	29.0	26.4
	Accent Adap + Speaker Adap (5 utt)	26.7	23.8
	Accent Adap + Speaker Adap (30 utt)	26.4	23.8

Table: Finnish Accented Speech, Word Error Rate %

Stacked Transformations

		WSJ-84	WSJCAM
Speech Recognition			
Speaker Adaptation			
Accented Speech	Baseline	40.6	32.1
	Speaker Adap (5 utt)	31.2	25.7
Accent and Neighbour Transformations	Speaker Adap (30 utt)	29.4	24.1
	5-Neighbour Adap	32.0	26.7
Stacked Transformations	5-Neighbour + Speaker Adap (5 utt)	28.5	24.5
Experiments	5-Neighbour + Speaker Adap (30 utt)	28.1	23.8
Conclusions	Accent Adap	29.0	26.4
	Accent Adap + Speaker Adap (5 utt)	26.7	23.8
	Accent Adap + Speaker Adap (30 utt)	26.4	23.8

Table: Finnish Accented Speech, Word Error Rate %

Stacked Transformations

		WSJ-84	WSJCAM
Speech Recognition			
Speaker Adaptation			
Accented Speech	Baseline	40.6	32.1
	Speaker Adap (5 utt)	31.2	25.7
Accent and Neighbour Transformations	Speaker Adap (30 utt)	29.4	24.1
	5-Neighbour Adap	32.0	26.7
Stacked Transformations	5-Neighbour + Speaker Adap (5 utt)	28.5	24.5
Experiments	5-Neighbour + Speaker Adap (30 utt)	28.1	23.8
Conclusions	Accent Adap	29.0	26.4
	Accent Adap + Speaker Adap (5 utt)	26.7	23.8
	Accent Adap + Speaker Adap (30 utt)	26.4	23.8

Table: Finnish Accented Speech, Word Error Rate %

Different amounts of Speaker Adaptation

Speech Recognition

Speaker Adaptation

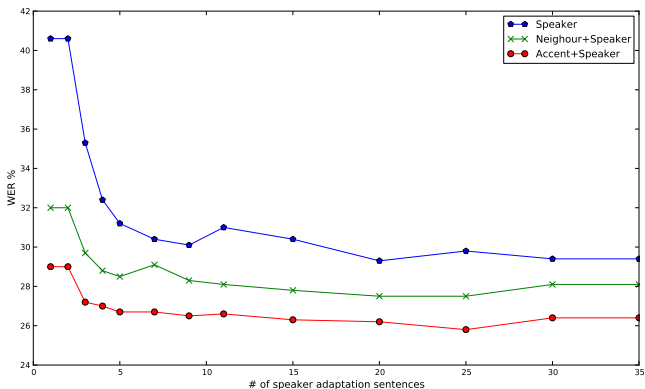
Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions



Conclusions

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- Both Neighbour and Accent Transformations give good results
- Stacked Transformations give better result than only Speaker Adaptation
- Stacked Transformations need less Speaker Adaptation data



Future - Use in Speech Synthesis

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- In Speech Synthesis it is hard to make a personalized voice with little adaptation data
- Possibly Accent Transformations can help to make accented voice
- Possibly Stacked Transformations can help to personalize a voice with less adaptation data



Questions?

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

■ Questions?



Questions?

Speech Recognition

Speaker Adaptation

Accented Speech

Accent and
Neighbour
Transformations

Stacked
Transformations

Experiments

Conclusions

- Questions?
- Slides and thesis (when final) can be found at users.ics.tkk.fi/peter

