Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction

M. Gibson[1], T. Hirsimäki[2], R. Karhila[2], M. Kurimo[2], W. Byrne[1]

[1]Cambridge University Engineering Department, [2]Helsinki University of Technology

EMIME Effective Multilingual Interaction in Mobile Environments

Introduction

Cross-lingual speaker adaptation:
- speaker data in *source* language
- HMM models in *target* language

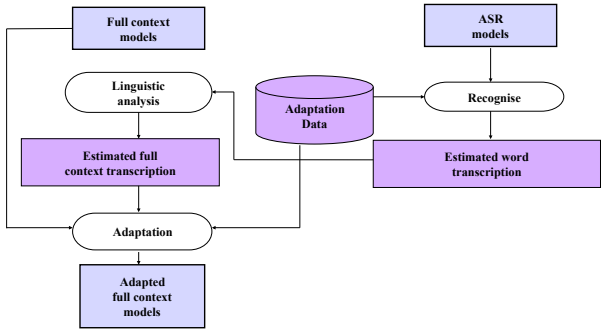
Previous approaches (see [1]):
- use source acoustic models
- learn and use mapping between source and target models
- no controlled comparison between intralingual and cross-lingual adaptation
- focus on supervised adaptation task

This work:
- uses no source language knowledge
- uses no mapping between source and target models
- controlled comparison with intralingual adaptation
- focusses on unsupervised adaptation task

Unsupervised adaptation (previous)

HMM-based synthesis uses *full context* models representing
- local phoneme context (quinphone)
- suprasegmental information e.g.
  - •syllabic stress, #syllables in word

Suprasegmental information renders full context models unsuitable for automatic speech recognition (ASR):
- adds complexity to recognition network construction

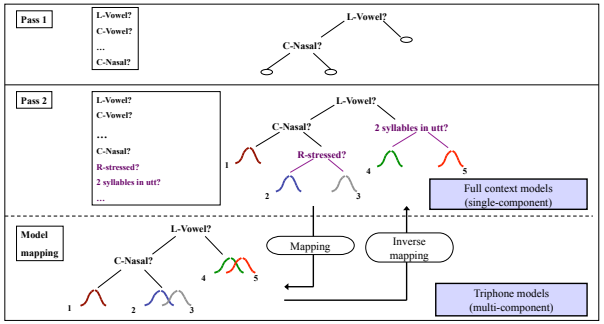


Unsupervised adaptation (issues)

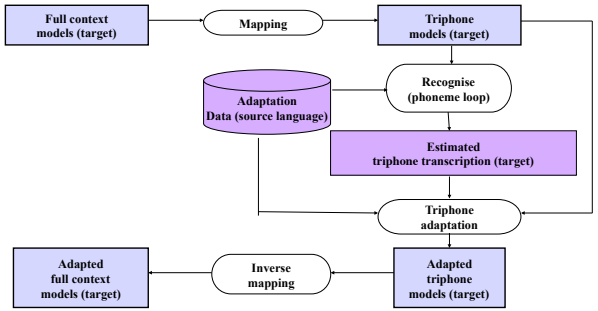
- Requires separately-trained ASR models
- Requires linguistic analysis of estimated transcription (but see [2] for an approach avoiding this)
- Lack of simple extension to cross-lingual case

Two-pass decision tree construction

- Defines mapping between single-component full context and multi-component triphone models (suitable for ASR)
- Mixture weights derived from occupancy of pass 2 leaf nodes



Unsupervised adaptation (two-pass)



- Source language same as target: unsupervised intralingual
- Source language differs from target: unsupervised cross-lingual

Evaluation

- English average voice (trained on SI84 WSJ)
- Adaptation data: 94 parallel translated utterances (native Finnish speaker) in
  - •Finnish (cross-lingual)
  - •English (intralingual comparison)
- 24 native English judges
- 1-5 rating of naturalness and similarity to target speaker

System	MOS naturalness	MOS similarity
Average voice	2.3	1.2
Adapted (unsupervised cross-lingual)	2.4	2.3
Adapted (unsupervised intralingual)	2.6	2.6
Adapted (supervised intralingual)	2.5	2.7
Vocoded natural speech	3.7	4.6

- No significant difference observed between naturalness or similarity of adapted systems

Conclusions

Proposed *unsupervised cross-lingual* adaptation method achieves similarity approaching that of *unsupervised intralingual* and *supervised intralingual* adaptation.

This is achieved without:
- **separately trained ASR models**
- **linguistic analysis of the estimated transcription**
- **knowledge of the source language of the adaptation data.**

Acknowledgements

This research was funded by the European Community's Seventh Framework Programme (FP7/2007-2013), grant agreement 213845 (EMIME).

References

[1] Y.Wu, Y.Nankaku, K.Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis", Interspeech 2009
[2] S.King, K. Tokuda, H. Zen, J. Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis", Interspeech 2008