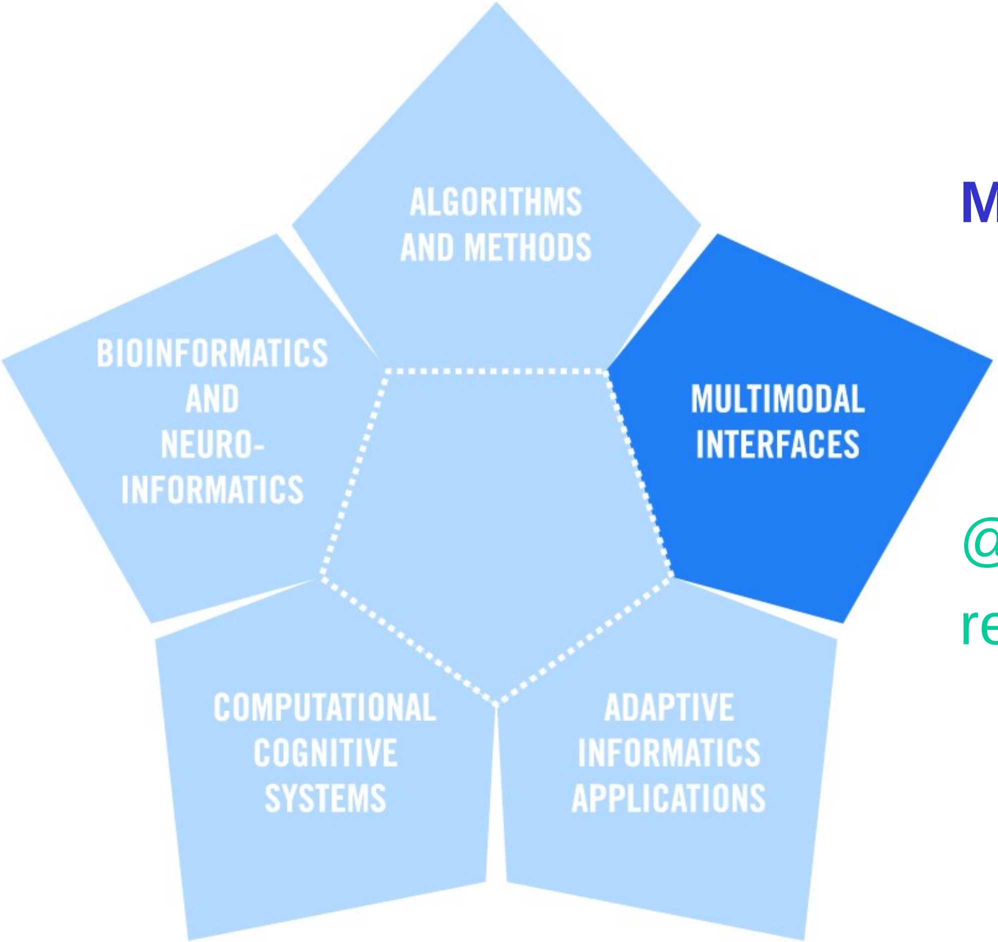
Aalto University

Context in multilingual speech processing – Adaptation of speech models

Mikko Kurimo, Reima Karhila, Peter Smit, Andre Mansikkaniemi

Adaptive Informatics Research Centre

Department of Information and Computer Science



Multimodal Interfaces group:

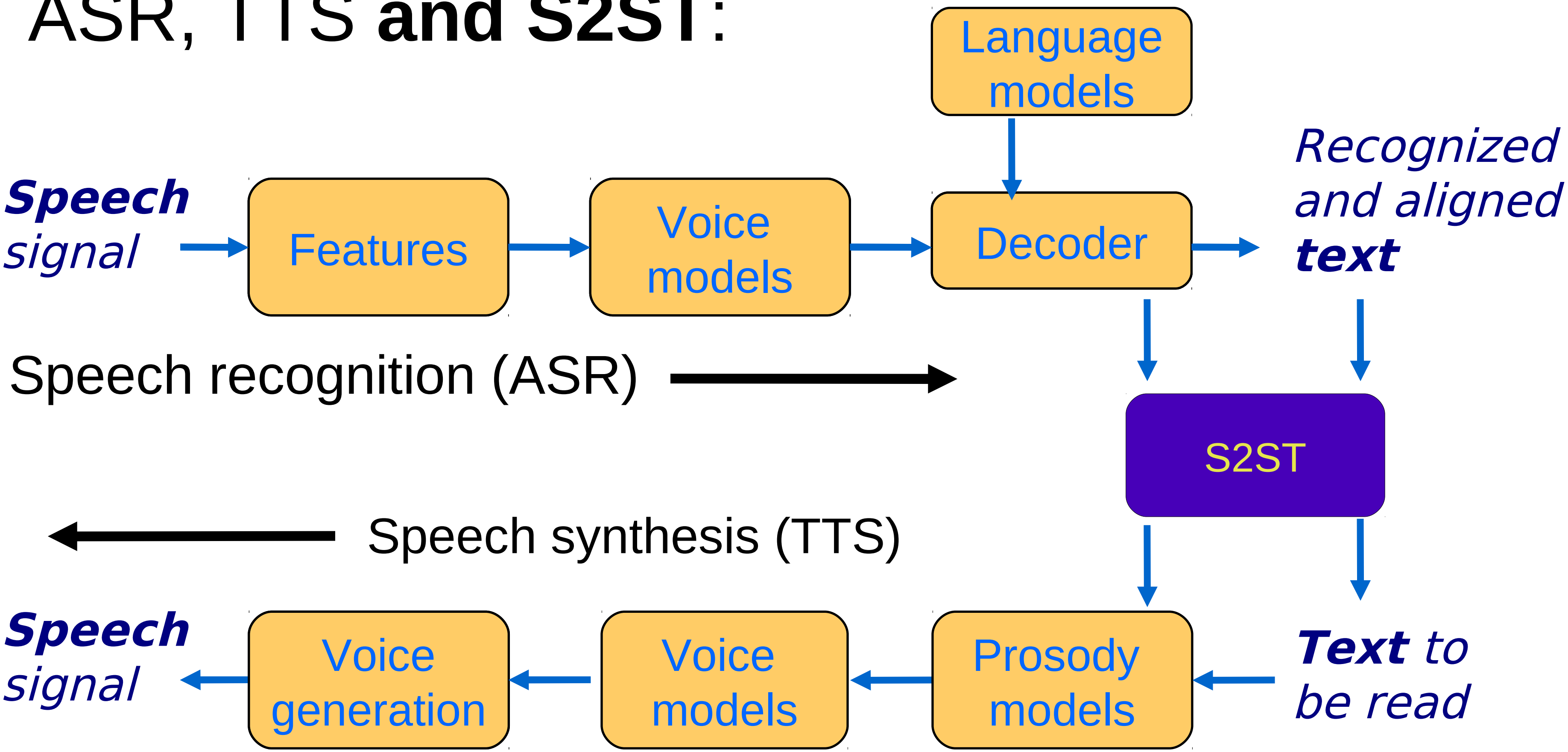
- Speech
- Language
- video & image

@Adaptive Informatics research centre (AIRC)

Contents

1. Why adapt voices to context?
2. Adaptation of speech models
   - for Automatic Speech Recognition (ASR)
   - for Text-To-Speech (TTS) synthesis
3. An example application demo:
   - Speech-To-Speech Translation (S2ST)

ASR, TTS **and S2ST**:



Applications for adapting speech models according to context

Speech recognition:

Dictation

Translation: input

Interfaces: input

Speech retrieval

Speech synthesis:

- Text reading
- Translation: output
- Interfaces: output
- Storing your personal voice

Personalized voice model: Model adapts to the speaker, speaking style, and environment

Why adapt voices to context?

An average voice (a statistical speech model trained using data from many speakers) performs quite well in ASR and TTS, but...

- Recognition accuracy degrades in noisy conditions
  - both for machine (ASR) and man (TTS)
- Poor ASR accuracy for non-standard speakers and styles
  - foreign accents, children, emotions, spontaneous etc.
- Personalization needed also for TTS
  - Speaker and speaking style (loudness, rate etc.)

Context information at various levels

Long span (wider than sentence)

- **Who speaks**, where, to whom?
- **Language, accent**, speaking style
- Topic, previous sentences, novelty
- Recording, background and acoustic environment

Short span (shorter than sentence)

- Previous words and sounds (language modeling)
- Pronunciation modeling (word and syllables)
- **Phoneme modeling** (neighboring phonemes)
- Acoustic features (spectral information)

An example of adaptation to context:
- ASR of foreign accented speech

Problem:

- The statistical speech models are trained for native speakers. Poor match for non-native accents.
- Not enough accent-specific training data for new models
- No time to collect enough adaptation data during recognition

Solution1: Stacked transformations*[Smit and Kurimo, ICASSP 2011]*

1. Adapt the native models by **accent-specific data** using maximum likelihood linear regression (MLLR)
2. Adapt the result further by **speaker-specific data** using another MLLR with less regression classes

Solution2: Cross-lingual speaker adaptation*[Karhila and Kurimo,SLT 2010]*

- Adapt the native models by speaker-specific **data in the foreign language** by first mapping the data to the other language (CLSA)

Stacked Transformations: Use data from similar speakers!

By accent

- Take accent-specific instead of speaker-specific data
- Data from a large number of speakers can be utilized
- Can be estimated before recognition

By speaker similarity

- Find several similar speakers (neighbours) using, e.g. eigenvoice parameters
- Use neighbours' data for adaptation
- Most computations can be done before recognition

Stacked Transformations

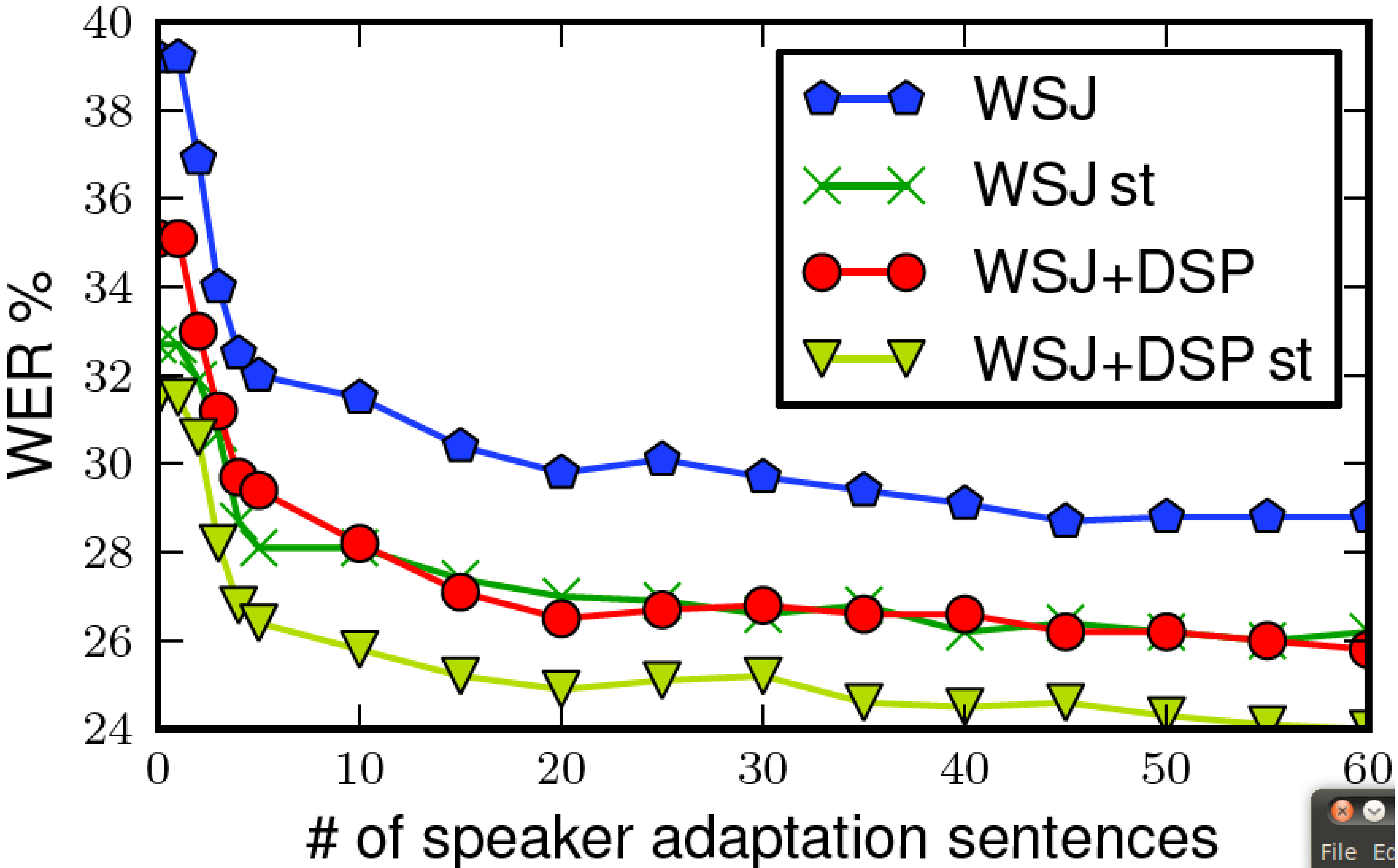
- Use first an Accent or Neighbour Transformation
  - More adaptation data is available, so more Regression Classes can be taken for the transformation
  - Supervised offline adaptation (transcription for the speech exists)
- Apply normal Speaker Adaptation after that
  - Less adaptation data, less Regression Classes
  - Unsupervised online adaptation (automatic transcription)
- Advantages
  - No cost increase at the recognition stage
  - The first transformation gives improved fit, so less speaker adaptation needed

Accent transformation experiments for ASR

	WSJ	*DSP*	*WSJ+DSP*	*WSJ at*	*WSJ+DSP at*
WSJ0	3.4	32.9	3.7	3.6	4.1
UED_Native	9.0	43.8	8.6	8.0	7.9
DSP	49.6	36.0	41.9	37.7	31.9
UED_Finnish	39.2	43.7	35.1	32.7	31.5

- Huge increase in WER for foreign-accented speech (Finnish)
- Not enough accented training data (DSP) for proper models
- Pooling all training data (WSJ + DSP) decreases WER significantly
- Accent transformation (at) helps even more

Foreign-accented speaker adaptation experiments for ASR



Solution2: Cross-lingual speaker adaptation (CLSA): Data from the same speaker in another language!

- **Unsupervised cross-lingual speaker adaptation** *[FP7 project EMIME 2008-2011]*
  - Adapt the native models by speaker-specific **data in the foreign language** by first mapping the data to the other language
    - E.g. to adapt English models using speaker characteristics learned from Finnish
  - Based on unsupervised transcription by cross-lingual ASR models
    - E.g. to recognize Finnish speech with English models
- **TTS experiment:** *[Karhila and Wester,Interspeech 2011]*
  - Adapt an English average voice by Finnish samples
  - Listen the result after 5, 15, 105 adaptation sentences
  - Compare native vs. foreign accented average voice

Mobile cross-lingual speech-to-speech interface

<u>EMIME 2008-2011:</u>

- Univ. Edinburgh
- Univ. Cambridge
- Nagoya Inst. Tech.
- IDIAP
- Nokia Res. Center
- Aalto University

<u>Features:</u>

- Speaker adaptation
- Cross-lingual
- Unified models for ASR and synthesis
- Mobile interface (N97)
- Fin, Eng, Chn, Jpn

Speech-to-Speech translation demo -Speaker adaptation for ASR and TTS

- **Real-time** version for travel phrases
- Run **ASR, MT, and TTS** on a server
- Server is typically a laptop connected by **WLAN**
- Interface for a **mobile phone** (N97)
- Versions for Finnish-English and Mandarin-English
- Using **speaker adaptation** to make the synthetic output speech sound more like the original speaker
- Adaptation is based on the unified HMM-based framework for ASR and TTS *[FP7 project EMIME 2008-2011]*

Applications of adapting speech models according to context

Speech recognition:

Dictation

Translation: input

Interfaces: input

Speech retrieval

Speech synthesis:

- Text reading
- Translation: output
- Interfaces: output
- Storing your personal voice

Personalized voice model: Model adapts to the speaker, speaking style, and environment