

USING STACKED TRANSFORMATIONS FOR RECOGNIZING FOREIGN ACCENTED SPEECH

Introduction

A common problem in foreign accented speech recognition is the lack of enough training data. In addition, training a model for a specific accent is computationally expensive. With an Accent Transformation, a generic model can be adapted to an Accent Dependent model. In addition, a Speaker Transformation could be 'stacked' on top of it. Experiments are performed for Accented and Stacked transformations with American English and Finnish accented English corpora.

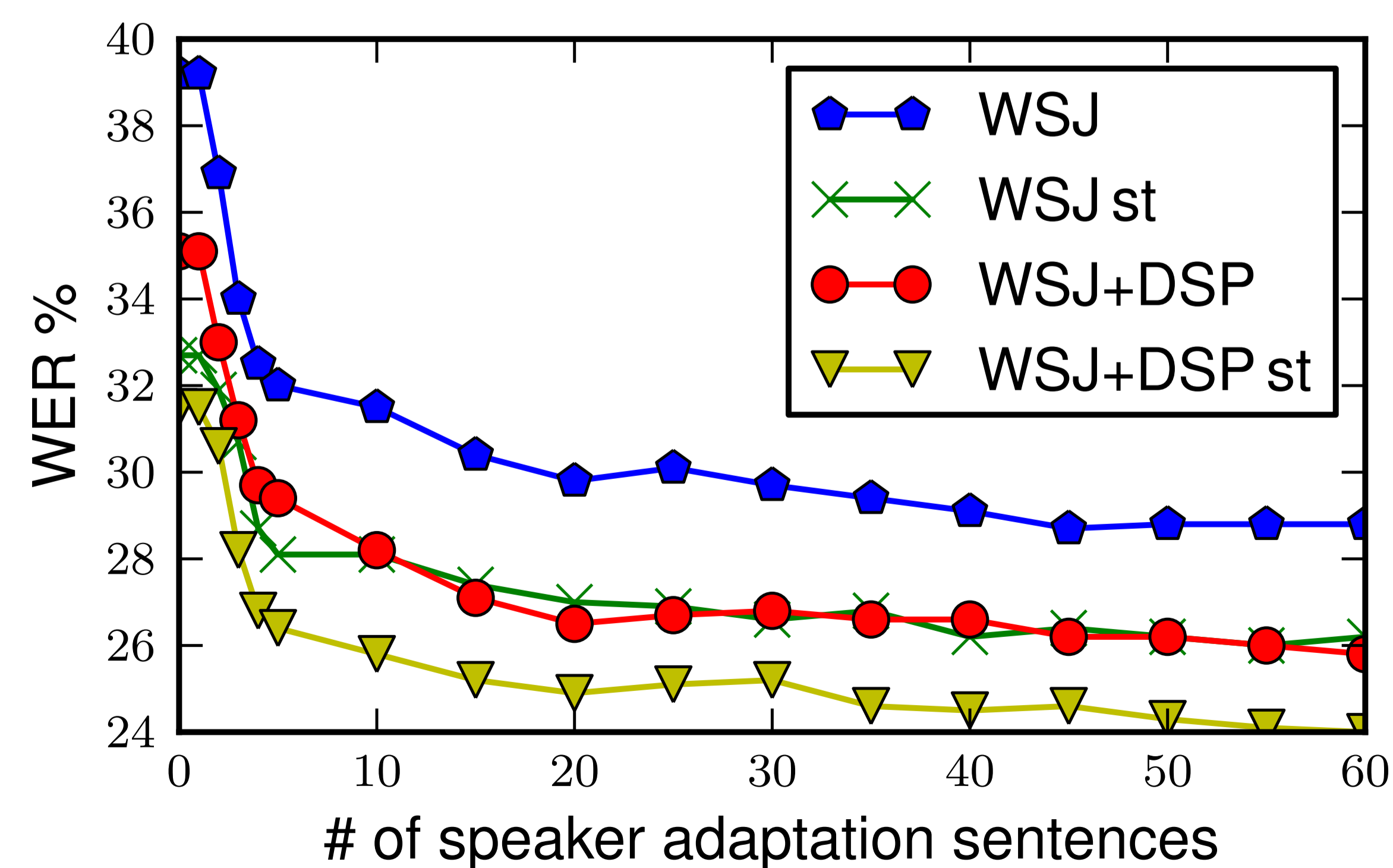
Experiments

As native corpora Wall Street Journal (WSJ) and a dataset from University of Edinburgh (UED_N) are used. Also two Finnish accented datasets are used, DSP and UED_F. In the experiment 'at' means Accent Transformation and 'st' Stacked Transformations. All results are in Word Error Rate %

Table: Accent Transformation evaluation

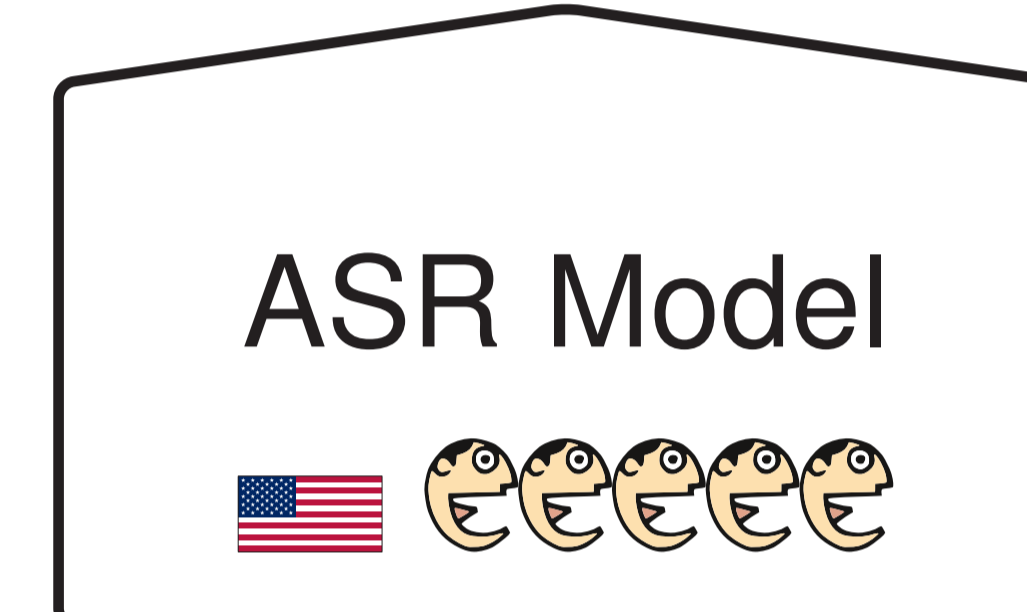
	WSJ	DSP	WSJ+DSP	WSJ at	WSJ+DSP at
WSJ0	3.4	32.9	3.7	3.6	4.1
UED_Native	9.0	43.8	8.6	8.0	7.9
DSP	49.6	36.0	41.9	37.7	31.9
UED_Finnish	39.2	43.7	35.1	32.7	31.5

Figure: Stacked Transformations evaluation



Transformation Schemes

- ▶ Speaker Transformation by Regression Class Tree CMLLR
- ▶ **Small amount** of adaptation data => Small number of Regression Classes
- ▶ Adaptation is often unsupervised (transcribed by ASR)
- ▶ Recognition needs **two passes**: first pass for transcription and second pass for recognition



- ▶ Accent Transformation by Regression Class Tree CMLLR
- ▶ Include **accented data** from different speakers
- ▶ **Large amount** of accented data => Large number of Regression Classes
- ▶ Adaptation is often supervised (transcription checked manually)
- ▶ Accent Transformations can be **calculated off line** (before recognition) => Single pass recognition



- ▶ Two transformations used in sequence; "**Stacked Transformations**"
- ▶ The detailed Accent Adaptation gives improved transcription for Speaker Adaptation
- ▶ Better match to the target speaker => Less Speaker Adaptation sentences needed
- ▶ Accent Transformation still calculated off line (before recognition)

Future Work

- ▶ Also other transformations are possible, e.g.
 - ▶ Neighbour Transformation – Choose some similar speakers, possibly in an unsupervised manner
 - ▶ Gender Transformation – Take speakers with same gender

Stacked Transformations in the EMIME project

This work was funded by the EMIME project. The EMIME project focused on developing personalized speech to speech translation. Stacked Transformations could be also used in HMM based synthesis. It could be used to create accented or personalized voices using less adaptation data.