

# ACL 2010 Demo Session

## Personalising a speech-to-speech translation in the EMIME project

Mikko Kurimo<sup>1</sup>, William Byrne<sup>6</sup>, John Dines<sup>3</sup>, Philip N. Garner<sup>3</sup>, Matthew Gibson<sup>6</sup>,

Yong Guan<sup>5</sup>, Teemu Hirsimäki<sup>1</sup>, Reima Karhila<sup>1</sup>, Simon King<sup>2</sup>, Hui Liang<sup>3</sup>,

Keiichiro Oura<sup>4</sup>, Lakshmi Saheer<sup>3</sup>, Matt Shannon<sup>6</sup>, Sayaka Shiota<sup>4</sup>, Jilei Tian<sup>5</sup>,

Keiichi Tokuda<sup>4</sup>, Mirjam Wester<sup>2</sup>, Yi-Jian Wu<sup>4</sup>, Junichi Yamagishi<sup>2</sup>

<sup>1</sup> Aalto University, Finland, <sup>2</sup> University of Edinburgh, UK,

<sup>3</sup> Idiap Research Institute, Switzerland, <sup>4</sup> Nagoya Institute of Technology, Japan,

<sup>5</sup> Nokia Research Center Beijing, China, <sup>6</sup> University of Cambridge, UK

<sup>7</sup> Corresponding author: [Mikko.Kurimo@tkk.fi](mailto:Mikko.Kurimo@tkk.fi)

### Abstract

In the EMIME project we have studied unsupervised cross-lingual speaker adaptation. We have employed an HMM statistical framework for both speech recognition and synthesis which provides transformation mechanisms to adapt the synthesized voice in TTS (text-to-speech) using the recognized voice in ASR (automatic speech recognition). An important application for this research is personalised speech-to-speech translation that will use the voice of the speaker in the input language to utter the translated sentences in the output language. In mobile environments this enhances the users' interaction across language barriers by making the output speech sound more like the original speaker's way of speaking, even she or he could not speak the output language.

Abstract length max 200 words

Paper length max 6 pages (for demos)

Submission DL 22 Feb (midnight)

## 1 Introduction

A mobile real-time speech-to-speech translation (STST) device is one of the grand challenges in natural language processing (NLP). It involves several important NLP research areas: automatic speech recognition (ASR), statistical machine translation (SMT) and speech synthesis, also known as text-to-speech (TTS). In recent years significant advance has also been made in relevant technological devices: the size of powerful computers has decreased to fit in a mobile phone and fast wifi and 3G networks have spread widely to connect them to even more powerful computation servers. Several hand-held STST applications and devices have become or are soon becoming available, for example (?; ?; ?; ?), but there are still

serious limitations in vocabulary and language selection and performance.

When an STST device is used in practical human interaction across a language barrier, one feature that is often missed is the personalization of the output voice. Who ever speaks to the device in what ever manner, the output voice sounds always the same. Producing high-quality synthesis voices is expensive and even if the system had many output voices, it is hard to select one that would sound like the input voice. There are many features in the output voice that could rise the interaction experience to a much more natural level, for example, emotions, speaking rate, loudness and the speaker identity.

After the recent development in hidden Markov model (HMM) based TTS, it has become possible to adapt the output voice using model transformations that can be estimated from small amount of speech samples. These techniques, like the maximum likelihood linear regression (MLLR), are adopted from HMM-based ASR where they are very powerful in fast adaptation of speaker and recording environment characteristics (?). Using hierarchical regression trees, the TTS and ASR models can further be coupled in a way that enables the unsupervised TTS adaptation (King et al., 2008). The unsupervised adaptation means that the unannotated adaptation samples can be sufficiently well annotated by ASR, and the voice transformation required for the TTS could be learned by ASR from the user's input speech in almost real-time.

The target in the EMIME project<sup>1</sup> is to study unsupervised cross-lingual speaker adaptation for STST systems. The first results of the project have been, for example, to bridge the gap between the ASR and TTS (?), to improve the baseline ASR (?) and SMT (de Gispert et al., 2009) systems for mor-

<sup>1</sup><http://emime.org>

phologically rich languages, and to develop the robust TTS (?). The next step has been the preliminary experiments in intra-lingual and cross-lingual speaker adaptation (?). For the cross-lingual adaptation several new methods have been proposed for mapping the HMM states (?), adaptation data (?) and model transformations (?).

In this presentation we can demonstrate the various new results in ASR, SMT and TTS. Even though the project is still going on, we have an initial version of the mobile STST system that will be used to show real-time cross-lingual speaker adaptation when the project is finished.

## 2 Baseline ASR, TTS and SMT systems

The baseline ASR systems in the project are developed using the HTK toolkit (?) for Finnish, English, Mandarin and Japanese. The systems can also utilize various real-time decoders such as Julius (?), Juicer (?) and the TKK decoder (?). The main structure of the baseline systems for each of the four languages is similar and fairly standard and in line with most other state-of-the-art large vocabulary ASR systems. Some special flavors have been added, such as the morphological analysis for Finnish (?). For speaker adaptation, the MLLR transformation based on hierarchical regression classes is included for all languages.

The baseline TTS systems in the project utilize the HTS toolkit (?) which is built on top of the HTK framework. The HMM-based TTS systems have been developed for Finnish, English, Mandarin and Japanese. The systems include an average voice model for each language trained over hundreds of speakers taken from standard ASR corpora, such as Specon (?). By the speaker adaptation transforms, thousands of new voices have been created (?) and new voices can be added using small amount of either supervised or unsupervised speech samples.

FIXME: Say something about the cross-lingual adaptation methods

Because the resources of the EMIME project have been focused on ASR, TTS and speaker adaptation, we aim at relying on the existing solutions for SMT as far as possible. New methods have been studied concerning the morphologically rich languages (de Gispert et al., 2009), but for the STST system we have are currently using the Google translate (?).

## 3 Demonstrations to show

### 3.1 Monolingual systems

In robust speech synthesis, a computer can learn to speak in the desired way 's after listening only a relatively small amount of training speech. The training speech can even be a normal quality recording outside the studio environment, where the target speaker is speaking to a standard microphone and the speech is not annotated. This differs dramatically from the conventional speech synthesis, where building a new voice required an hour or more careful repetition of specially selected prompts recorded in an anechoic chamber with high quality equipment.

The robust speech synthesis has recently become possible using the statistical HMM framework for both ASR and TTS. This framework enables the use of efficient speaker adaptation transformations developed for ASR to be used also for the TTS models. Using the large corpora collected for ASR, we can train average voice models for both ASR and TTS. The training data may include a small amount of speech with poor coverage of phonetic contexts from each single speaker, but by summing the material over hundreds of speakers, we can get sufficient models for an average speaker. Only a small amount of adaptation data is then required to create transformations for tuning the average voice closer to the target voice.

In addition to the supervised adaptation using annotated speech, it is also possible to create sufficient annotations by ASR from unannotated speech. This unsupervised adaptation enables the system to use a much broader selection of sources, for example, recorded samples in the internet, to learn a new voice.

The following systems can demonstrate the results of monolingual adaptation:

1. In *EMIME Voice cloning in Finnish and English* the goal is that the users can clone their own voice. The user will dictate for about 10 minutes and then after half an hour of processing time, the TTS system has transformed the average model towards the user's voice and can speak with this voice. The cloned voices may become especially valuable, for example, if a person's voice is later damaged in an accident or by a disease.
2. In *EMIME Thousand voices map* the goal is to browse the world's largest collection of

synthetic voices by using a world map interface (?). The user can zoom in the world map and select any voice, which are organized according to the place of living of the adapted speaker, to utter the given sentence.

3. The models developed in the HMM framework can be demonstrated also in adaptation of an ASR system for *large-vocabulary continuous speech recognition*. By utilizing morpheme-based language models instead of word-based models the Finnish ASR system is able to cover practically an unlimited vocabulary (?). This is necessary for morphologically rich languages where, due to inflection, derivation and composition, there exists so many different word forms that word based language modeling becomes impractical.

### 3.2 Cross-lingual systems

In the EMIME project the goal is to learn cross-lingual speaker adaptation. Here the output language ASR or TTS system is adapted from speech samples in the input language. The results so far are encouraging, especially for TTS: Even though the cross-lingual adaptation may somewhat degrade the synthesis quality, the adapted speech sounds now more like the target speaker (?).

The following systems have been created to demonstrate cross-lingual adaptation:

1. In *EMIME Cross-lingual Finnish/English and Mandarin/English TTS adaptation* the input languages sentences dictated by the user will be used to learn the characteristics of her or his voice. The adapted cross-lingual model will then be used to speak the given output language sentences in the same user's voice in output language. It is not required that the user is bilingual or speaks or understands the output language as all the required spoken input is in his own native language.

FIXME: Could the PC voice cloning be extended to this?

2. In *EMIME Real-time speech-to-speech mobile translation demo* two users will interact using a pair of mobile N97 devices. The system will recognize the phrase the other user is speaking in his native language and translate and speak it in the native language of the

other user. After a few sentences the system will have the speaker adaptation transformations ready and can apply them in the synthesized voices to make them sound more like the original speaker instead of a standard voice. The first real-time demo will be available for the Mandarin/English language pair.

FIXME: Does this work enough that we dare to try a public demo?

FIXME: Get pics and text from D4.4. / Nokia

3. *The morpheme-based translation system* for Finnish/English and English/Finnish can be compared to a word based translation for arbitrary sentences. The morpheme-based approach is particularly useful for language pairs where one or both languages are morphologically rich ones where the amount and complexity of different word forms severely limits the performance for word-based translation. The morpheme-based systems can learn translation models for phrases where morphemes are used instead of words (de Gispert et al., 2009).

## 4 Future challenges

### Acknowledgments

### References

- Adria de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. 2009. Minimum bayes risk combination of translation hypothesis from alternative morphological decompositions. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the association for Computational Linguistics, NAACL 2009*, Boulder, CO, May 31 - June 5.
- S. King, K. Tokuda, H. Zen, and J. Yamagishi. 2008. Unsupervised adaptation for HMM-based speech synthesis. In *Proc. Interspeech 2008*, pages 1869–1872, September.