Personalising speech-to-speech translation in the EMIME project

Mikko Kurimo[1†], William Byrne[6], John Dines[3], Philip N. Garner[3], Matthew Gibson[6], Yong Guan[5], Teemu Hirsimäki[1], Reima Karhila[1], Simon King[2], Hui Liang[3], Keiichiro Oura[4], Lakshmi Saheer[3], Matt Shannon[6], Sayaka Shiota[4], Jilei Tian[5], Keiichi Tokuda[4], Mirjam Wester[2], Yi-Jian Wu[4], Junichi Yamagishi[2]

[1] Aalto University, Finland, [2] University of Edinburgh, UK, [3] Idiap Research Institute, Switzerland, [4] Nagoya Institute of Technology, Japan, [5] Nokia Research Center Beijing, China, [6] University of Cambridge, UK

[†]Corresponding author: 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 if she or he could not speak the output language.

1 Introduction

A mobile real-time speech-to-speech translation (S2ST) 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 have 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 S2ST applications and devices have already become available, for example by IBM, Google or Jibbigo[1], but there are still serious limitations in vocabulary and language selection and performance.

When an S2ST device is used in practical human interaction across a language barrier, one feature that is often missed is the personalization of the output voice. Whoever speaks to the device in what ever manner, the output voice always sounds 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 raise 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 a small number of speech samples. These techniques, for instance 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 (Gales, 1998). Using hierarchical regression trees, the TTS and ASR models can further be coupled in a way that enables unsupervised TTS adaptation (King et al., 2008). In unsupervised adaptation samples are annotated by applying ASR. By eliminating the need for human intervention it becomes possible to perform voice adaptation for TTS in almost real-time.

The target in the EMIME project[2] is to study unsupervised cross-lingual speaker adaptation for S2ST systems. The first results of the project have

[1] http://www.jibbigo.com
[2] http://emime.org

Proceedings of the ACL 2010 System Demonstrations, pages 48–53,
Uppsala, Sweden, 13 July 2010.

been, for example, to bridge the gap between the ASR and TTS (Dines et al., 2009), to improve the baseline ASR (Hirsimäki et al., 2009) and SMT (de Gispert et al., 2009) systems for morphologically rich languages, and to develop robust TTS (Yamagishi et al., 2010). The next step has been preliminary experiments in intra-lingual and cross-lingual speaker adaptation (Wu et al., 2008). For cross-lingual adaptation several new methods have been proposed for mapping the HMM states, adaptation data and model transformations (Wu et al., 2009).

In this presentation we can demonstrate the various new results in ASR, SMT and TTS. Even though the project is still ongoing, we have an initial version of mobile S2ST system and cross-lingual speaker adaptation to show.

2 Baseline ASR, TTS and SMT systems

The baseline ASR systems in the project are developed using the HTK toolkit (Young et al., 2001) for Finnish, English, Mandarin and Japanese. The systems can also utilize various real-time decoders such as Julius (Kawahara et al., 2000), Juicer at IDIAP and the TKK decoder (Hirsimäki et al., 2006). 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 for have been added, such as the morphological analysis for Finnish (Hirsimäki et al., 2009). 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 (Yamagishi et al., 2009) 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 Speecon (Iskra et al., 2002). Using speaker adaptation transforms, thousands of new voices have been created (Yamagishi et al., 2010) and new voices can be added using a small number of either supervised or unsupervised speech samples. Cross-lingual adaptation is possible by creating a mapping between the HMM states in the input and the output language (Wu et al., 2009).

Because the resources of the EMIME project have been focused on ASR, TTS and speaker adaptation, we aim at relying on 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 S2ST system we are currently using Google translate[3].

3 Demonstrations to show

3.1 Monolingual systems

In robust speech synthesis, a computer can learn to speak in the desired way after processing 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 conventional TTS, where building a new voice requires an hour or more careful repetition of specially selected prompts recorded in an anechoic chamber with high quality equipment.

Robust TTS 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 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 obtain 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 employ ASR to create annotations. This unsupervised adaptation enables the system to use a much broader selection of sources, for example, recorded samples from the internet, to learn a new voice.

The following systems will 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

[3] `http://translate.google.com`

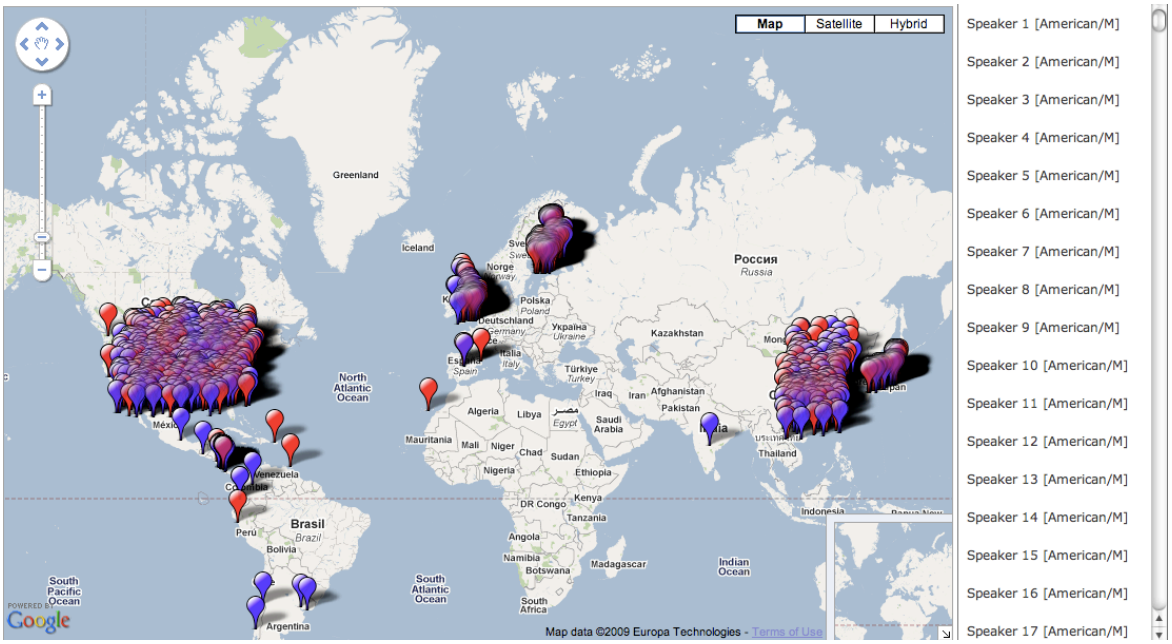

Figure 1: Geographical representation of HTS voices trained on ASR corpora for EMIME projects. Blue markers show male speakers and red markers show female speakers. Available online via http://www.emime.org/learn/speech-synthesis/listen/Examples-for-D2.1

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 (Yamagishi et al., 2010). 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. This interactive geographical representation is shown in Figure 1. Each marker corresponds to an individual speaker. Blue markers show male speakers and red markers show female speakers. Some markers are in arbitrary locations (in the correct country) because precise location information is not available for all speakers. This geographical representation, which includes an interactive TTS demonstration of many of the voices, is available from the URL provided. Clicking on a marker will play synthetic speech from that speaker[4]. As well as being a convenient interface to compare the many voices, the interactive map is an attractive and easy-to-understand demonstration of the technology being developed in EMIME.

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 (Hirsimäki et al., 2006). 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 now sounds more like the target speaker. Several recent evaluations of the cross-lingual speaker

[4]Currently the interactive mode supports English and Spanish only. For other languages this only provides pre-synthesised examples, but we plan to add an interactive type-in text-to-speech feature in the near future.

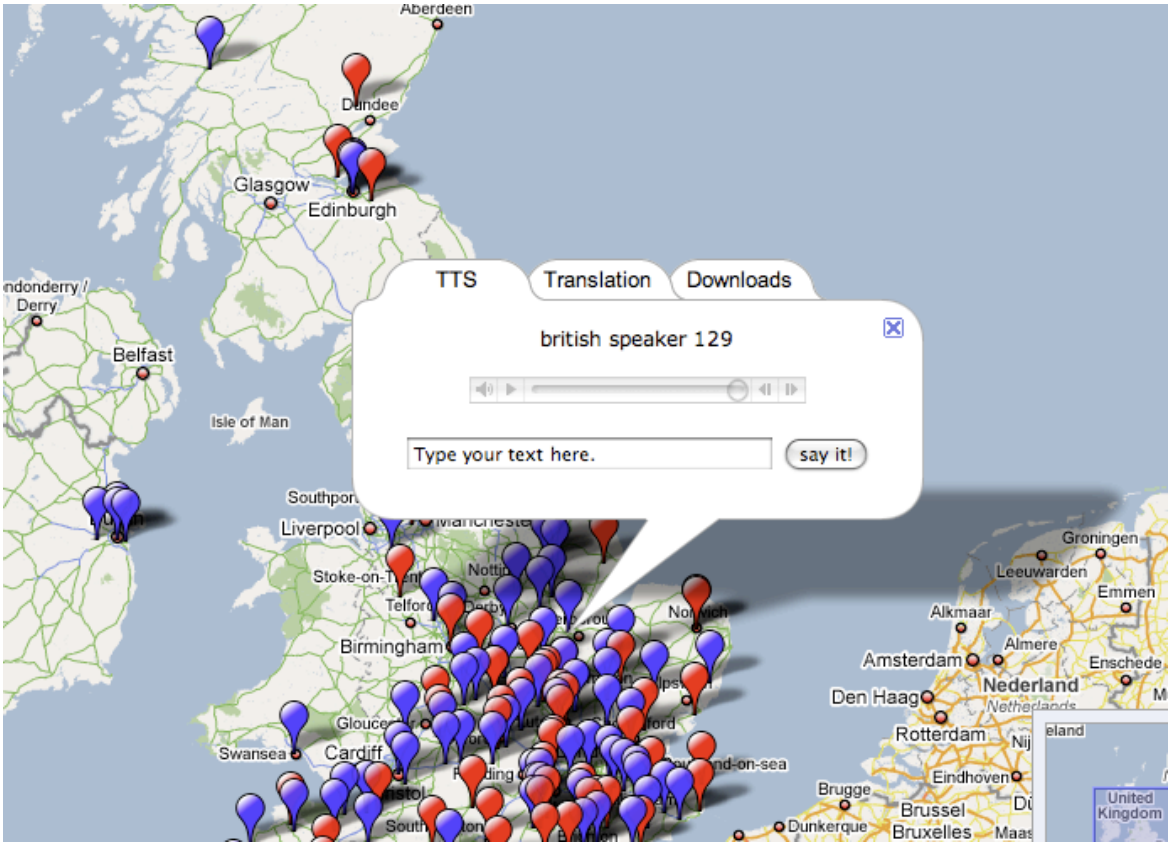

Figure 2: All English HTS voices can be used as online TTS on the geographical map.

adaptation methods can be found in (Gibson et al., 2010; Oura et al., 2010; Liang et al., 2010; Oura et al., 2009).

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 language sentences dictated by the user will be used to learn the characteristics of her or his voice. The adapted cross-lingual model will be used to speak output language (English) sentences in the user's voice. The user does not need to be bilingual and only reads sentences in their native language.

2. In *EMIME Real-time speech-to-speech mobile translation demo* two users will interact using a pair of mobile N97 devices (see Figure 3). 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 version is available for the Mandarin/English language pair.

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). Recent evaluations (Kurimo et al., 2009) have shown that the performance of the unsupervised data-driven morpheme segmentation can rival the conventional rule-based ones. This is very useful if hand-crafted morphological analyzers are not available or their coverage is not sufficient for all languages.

Acknowledgments

The research leading to these results was partly funded from the European Communitys Seventh

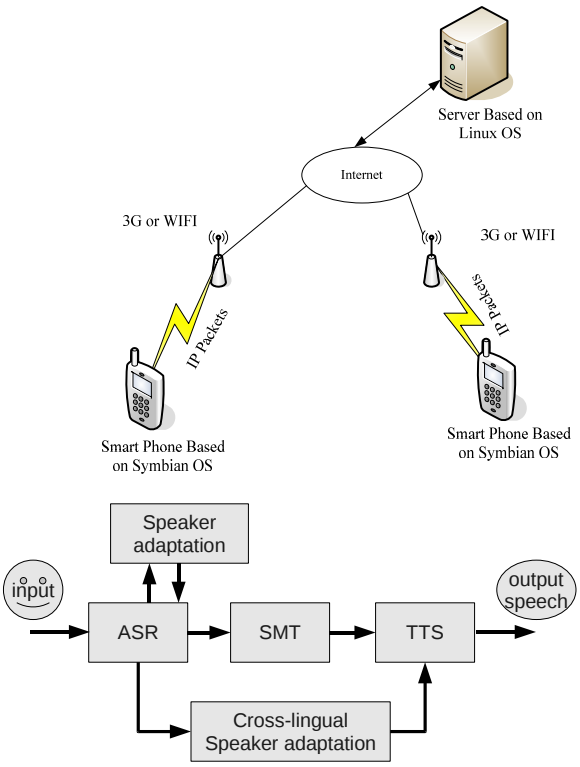

Figure 3: EMIME Real-time speech-to-speech mobile translation demo

Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

References

A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne. 2009. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proc. NAACL-HLT*.

J. Dines, J. Yamagishi, and S. King. 2009. Measuring the gap between HMM-based ASR and TTS. In *Proc. Interspeech '09*, Brighton, UK.

M. Gales. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12(2):75–98.

M. Gibson, T. Hirsimäki, R. Karhila, M. Kurimo, and W. Byrne. 2010. Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction. In *Proc. of ICASSP*, page to appear, March.

T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to finnish. *Computer Speech & Language*, 20(4):515–541, October.

T. Hirsimäki, J. Pylkkönen, and M Kurimo. 2009. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Trans. Audio, Speech, and Language Process.*, 17:724–732.

D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling. 2002. SPEECON speech databases for consumer devices: Database specification and validation. In *Proc. LREC*, pages 329–333.

T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano. 2000. Free software toolkit for japanese large vocabulary continuous speech recognition. In *Proc. ICSLP-2000*, volume 4, pages 476–479.

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.

Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. 2009. Overview and results of Morpho Challenge 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September.

H. Liang, J. Dines, and L. Saheer. 2010. A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis. In *Proc. of ICASSP*, page to appear, March.

Keiichiro Oura, Junichi Yamagishi, Simon King, Mirjam Wester, and Keiichi Tokuda. 2009. Unsupervised speaker adaptation for speech-to-speech translation system. In *Proc. SLP (Spoken Language Processing)*, number 356 in 109, pages 13–18.

K. Oura, K. Tokuda, J. Yamagishi, S. King, and M. Wester. 2010. Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis. In *Proc. of ICASSP*, page to appear, March.

Y.-J. Wu, S. King, and K. Tokuda. 2008. Cross-lingual speaker adaptation for HMM-based speech synthesis. In *Proc. of ISCSLP*, pages 1–4, December.

Y.-J. Wu, Y. Nankaku, and K. Tokuda. 2009. State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In *Proc. of Interspeech*, pages 528–531, September.

J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals. 2009. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Trans. Audio, Speech and Language Process.*, 17(6):1208–1230. (in press).

J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo. 2010. Thousands of voices for hmm-based speech synthesis. *IEEE Trans. Speech, Audio & Language Process.* (in press).

S. Young, G. Everman, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, 2001. *The HTK Book Version 3.1*, December.