www.emime.org

Effective Multilingual Interaction in Mobile Environments

"The last barrier for global e-commerce is the language barrier."
Robert Levin, Chief Executive TransClick, Forbes.com

EMIME will help to overcome the language barrier by developing a mobile device that performs personalised speech-to-speech translation, such that the a user's spoken input in one language is used to produce spoken output in another language while continuing to sound like the user's voice.

The tourism industry is the single largest contributor to global gross domestic product, and business operations are frequently carried out across geographical and linguistic borders. Yet, cross-linguality means that one of the most elementary and crucial elements of human communication – spoken language – remains a fundamental barrier to progress. It is clear that a key to breaking down this language barrier is computer-assisted interaction, but the ideal solution of a portable 'universal translator', in which cross-lingual spoken interaction is instantaneously and seamlessly facilitated by an unobtrusive automated assistant, still remains only a vision for the future. Even so, the critical elements that would comprise such a system – automatic speech recognition (ASR), machine translation and text-to-speech synthesis (TTS) – have made dramatic leaps in performance in the last decade and progress in these fields will continue to bring such a device closer to reality.

As a result of these advances, several research and commercially-based speech-to-speech translation efforts have been made in recent years – to mention only a few: *Verbmobil* a long-term project of the German Federal Ministry of Education, Science, Research and Technology, *Technology and Corpora for Speech to Speech Translation* (TC-STAR) FP6 European project, and the *Global Autonomous Language Exploitation* (GALE) DARPA initiative. Ranging from constrained, mobile applications to ambitious systems demanding considerable computing power, these efforts demonstrate that there is a strong demand for such technology across a broad spectrum of applications. One aspect which we take for granted in spoken communication and that is missing from current technology is a means to facilitate the *personal* nature of spoken language. That is, state-of-the-art approaches lack the ability to be personalised in an effective and unobtrusive manner. This lack of personalisation is a significant barrier to natural communication.

In EMIME we are extending the state-of-the-art, with an emphasis on *adaptive*, *personalised*, *cross-lingual* speech processing.

> "A major research challenge involves the shift from explicit interaction to implicit human-computer interaction, where the computer adapts to the user, their language and context. In addition, we must develop ways to create adaptive interfaces that enable personalised interaction."
> *Xavier Gros, DG INFSO, European Commission, LREC 2006 Workshop, Genova.*

Personalisation of systems for cross-lingual spoken communication is an important, but little explored, topic. It would provide more natural interaction and make the computing device a less obtrusive, but still essential element in assisting such human-human interactions. Research in this area poses new technological challenges and will open up exciting new possibilities in the development of such systems. In particular, it will call on the development of unified approaches for the modelling of speech for recognition and synthesis that will need to adapt across languages to each user's speaking characteristics. We believe that, within a restricted domain of limited lexical and grammatical complexity, it is now possible to develop techniques for speech-to-speech translation that can be personalised to the user and that such technology can form the basis of useful mobile devices for assisting cross-lingual spoken interactions.

The EMIME project concerns *intuitive multimodal interfaces and interpersonal communication systems* and aims to make a significant contribution towards the vision of an effective system for assisting cross-lingual interaction in realistic, constrained application scenarios. We will build a mobile device that carries out *personalised* speech-to-speech translation. **The users' spoken input in one language will be used to produce spoken output in a target language that will sound like that of the user.**

The project has five objectives:

Objective 1	We will personalise speech processing systems by learning individual characteristics of a user's speech and reproducing them in synthesised speech.
Objective 2	We will introduce a cross-lingual capability such that personal characteristics can be reproduced in a second language not spoken by the user.
Objective 3	We will develop and better understand the mathematical and theoretical relationship between speech recognition and synthesis.
Objective 4	We will eliminate the need for human intervention in the process of cross-lingual personalisation.
Objective 5	We will evaluate our research against state-of-the art techniques in component-wise and end-to-end systems and demonstrate our achievements in a practical mobile application.

The project will take three years and has just reached the end of the first year. We have been conducting a large series of experiments to identify the capabilities and limitations of current technology, for speech recognition and speech synthesis. We have experimented with novel configurations of speech recognisers that are closer to typical speech synthesis configurations, and vice versa, in a first attempt to 'bridge the gap' between the two.

The key outputs from the first year of the project are primarily project-internal tools and baseline systems that lay the foundations we need for the research in the next two years. But we have already made some important discoveries and invented new techniques. We have, for the first time, been able to create a virtually unlimited number of different voices for a speech synthesiser. Until now, creating a new voice was a laborious and expensive process – that is why most commercial systems have a very small set of voices from which to choose. Our method can create hundreds or thousands of voices using commonly-available databases of recorded speech. The technique can work with low-quality recordings and small amounts of speech. It can be applied to any accent of any language. To demonstrate what this technique is capable of, we created a graphical interface that allows you to explore the many voices we have built so far in EMIME and other projects (nearly 1000). You can access it on the web at `http://www.emime.org/learn/speech-synthesis/listen/voices-of-the-world`.