**Workpackage 5 : Deliverable D5.6**
**Project final report**


**Effective Multilingual Interaction in Mobile Environments**

| | |
|---|---|
| **Date of preparation** | April 2011 |
| **Version number** | 1 |
| **Grant Agreement number** | FP7-213845 |
| **Project acronym** | EMIME |
| **Project title** | Effective Multilingual Interaction in Mobile Environments |
| **Funding Scheme** | FP7 THEME ICT-2007.2.1 |
| **Date of latest version of Annex I against which the assessment will be made** | 31 Oct 2007 |
| **Periodic report** | 3rd |
| **Period covered** | 1 March 2010 to 28 February 2011 |
| **Project co-ordinator** | Prof. Simon King, University of Edinburgh |
| **Tel** | + 44 131 651 1725 |
| **Fax** | + 44 131 650 6626 |
| **Email** | Simon.King@ed.ac.uk |
| **Project website address** | `http://www.emime.org` |

| Participant no. | Participant organisation name | Part. short name | Country |
|---|---|---|---|
| 1 (Coordinator) | University of Edinburgh | UEDIN | UK |
| 2 | Idiap Research Institute | IDIAP | Switzerland |
| 3 | Aalto University | AALTO | Finland |
| 4 | Nagoya Institute of Technology | NIT | Japan |
| 5 | Nokia Corporation | NOKIA | Finland |
| 6 | University of Cambridge | UCAM | UK |

# Contents

# 4.1

# Final publishable summary report

This part of the report is broken down into the requisite five sections, with page breaks inserted to allow easy extraction of sections as appropriate.

### 4.1.1 Executive summary

One of the most elementary and crucial elements of human communication – spoken language – remains a fundamental barrier to progress because of the difficulty of communicating across languages. The 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.

The goal of the EMIME project was to enhance such speech-to-speech translation devices by producing personalised speech synthesis. In order to demonstrate this, we have created a speech-to-speech translation system in which the output translated speech sounds like the input speaker.

The key technology that we developed is a method for personalising speech synthesis within and across languages. This technology enables the fully automatic construction of an unlimited number of different synthetic voices, starting only from untranscribed recordings of speech. Since speech data are cheap and easy to obtain, the technology has a lot of potential in a much wider range of applications, such as

- creating personalised voices for assistive communication devices – for example, for people who may lose their voice through a degenerative condition
- social applications
- entertainment and games applications

The project lasted three years:

- In the first year, the project outputs were primarily project-internal tools and baseline systems that lay the foundations we need for the research in the next two years. But we already made some important discoveries and invented new techniques which allow us to make new synthetic voices quickly and automatically using small amounts of data. To demonstrate what this technique is capable of, we created a graphical interface at `http://www.emime.org/learn/speech-synthesis/listen/voices-of-the-world`.
- In the second year, we developed and compared a number of different methods for capturing speaker characteristics in one language and transferring them to another language. This is the main contribution of this project: cross-language speaker adaptation. It enables us to reach our goal of personalised speech translation. The third year will involve refinement of this methods and implementation of them within an end-to-end speech translation system.
- In the third year, we integrated and refined these methods and produced both a laboratory-based "research demonstrator" which showcases the highest quality version of our methods and a "realtime demonstrator" which demonstrates how the method can by implemented in a live running application. We conducted extensive evaluations to test the effectiveness of our methods and in the process also discovered more about human performance in identifying speakers across languages.

We have released the project foreground as open source software and data, enabling commercial exploitation and continued research.

### 4.1.2 Summary description of project context and objectives

> "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.*

One of the most elementary and crucial elements of human communication – spoken language – remains a fundamental barrier to progress because of the difficulty of communicating across languages. The 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.

The goal of the EMIME project was to enhance such speech-to-speech translation devices by producing personalised speech synthesis. In order to demonstrate this, we have created a speech-to-speech translation system in which the output translated speech sounds like the input speaker.

The key technology that we developed is a method for personalising speech synthesis within and across languages. This technology enables the fully automatic construction of an unlimited number of different synthetic voices, starting only from untranscribed recordings of speech. Since speech data are cheap and easy to obtain, the technology has a lot of potential in a much wider range of applications, such as

- creating personalised voices for assistive communication devices – for example, for people who may lose their voice through a degenerative condition
- social applications
- entertainment and games applications

Personalisation of systems for cross-lingual spoken communication is an important, but little explored, topic. It will 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 concerned *intuitive multimodal interfaces and interpersonal communication systems* and has make a significant contribution towards the vision of an effective system for assisting cross-lingual interaction in realistic, constrained application scenarios. We have built a mobile device that carries out *personalised* speech-to-speech translation. **The users' spoken input in one language is used to produce spoken output in a target language that sounds like that of the user.**

The project achieved the following five objectives:

***Objective 1*** To personalise speech processing systems by learning individual characteristics of a user's speech and reproducing them in synthesised speech.

***Objective 2*** To introduce a cross-lingual capability such that personal characteristics can be reproduced in a second language not spoken by the user.

***Objective 3*** To develop and better understand the mathematical and theoretical relationship between speech recognition and synthesis.

***Objective 4*** To eliminate the need for human intervention in the process of cross-lingual personalisation.

***Objective 5*** To 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 took three years and our key outputs were:

- In the first year, the project outputs were primarily project-internal tools and baseline systems that lay the foundations we need for the research in the next two years. But we already made some important discoveries and invented new techniques which allow us to make new synthetic voices quickly and automatically using small amounts of data. To demonstrate what this technique is capable of, we created a graphical interface at `http://www.emime.org/learn/speech-synthesis/listen/voices-of-the-world`.
- In the second year, we developed and compared a number of different methods for capturing speaker characteristics in one language and transferring them to another language. This is the main contribution of this project: cross-language speaker adaptation. It enables us to reach our goal of personalised speech translation. The third year will involve refinement of this methods and implementation of them within an end-to-end speech translation system.
- In the third year, we integrated and refined these methods and produced both a laboratory-based "research demonstrator" which showcases the highest quality version of our methods and a "realtime demonstrator" which demonstrates how the method can by implemented in a live running application. We conducted extensive evaluations to test the effectiveness of our methods and in the process also discovered more about human performance in identifying speakers across languages.

## 4.1.3 Main S&T results/foregrounds

We report the main science and technology results by workpackage, since these map on to the main themes of research of the EMIME project. The exploitable foreground, which is being publicly released as open source, is described in Section 4.2 – B.2.

### 4.1.3.1 WP1: Infrastructure – data, machine translation resources, application scenario definition

Workpackage 1 was relatively small and its main task was the collection of resources for use in the other workpackages. We are releasing the data collected in WP1.

#### 4.1.3.1.1 Overall objectives of WP1

- Identification of the languages to be studied.

  *Progress in Year 1: After a survey of available resources, and after considering the research interests and language expertise of the EMIME partners, we selected the following languages for study: English (American accented), Mandarin, Japanese and Finnish.*

- Identification and distribution of monolingual speech and text collections (in multiple languages) to be used in the development of baseline speech recognition and speech synthesis systems as well as the development of speaker adaptation procedures (see WP2).

  *Progress in Year 1: Monolingual speech and text collections (in multiple languages) were selected to be used in the development of baseline speech recognition and speech synthesis systems as well as the development of speaker adaptation procedures. These corpora are listed in D1.1. Thus far in the project we had used the following corpora for English: CMU-ARCTIC, RM, WSJ0, WSJ1, WSJCAM0 and a collection of UEDIN's speech synthesis voices. For Mandarin and Finnish, we had used Speecon data. For Japanese we had used CSJ and a collection of NIT's speech synthesis corpora. A common ftp server was used to share data.*

- Identification and distribution of multilingual speech and text collections to be used in the development of baseline speech recognition and speech synthesis systems as well as the development of speaker adaptation procedures (see WP3).

  *Progress in Year 2: The languages studied were clearly identified and the related resources made available.*

- Provision of automatic translation of reference transcriptions as well as recognition output for use in creating project-wide collections for development and evaluation of acoustic adaptation and speech synthesis.

  *Progress in Year 3: recognition and translation output, including N-best lists, were created for project-wide use*

#### 4.1.3.1.2 Significant results obtained in WP1

***Deliverable D1.1***

This deliverable (Identification of Languages and Resources for Study in Work Package 2), containing further details on the above tasks, was completed on time and successfully identified all the required resources for the work conducted in the project to date.

***Software resources***

Definitive versions of the various software to be used in the project were identified. Notably, the HTS version of HTK was selected for use in both TTS and ASR experiments. The choice of decoder was initially deferred since two of the main contenders (Julius and Juicer) were both under active development during the project to support various features of synthesis-type HMMs. Finally, we continued to use both decoders.

***Common data sets for ASR and TTS***

In anticipation of experiments conducted in WP2, D1.1 suggested the use of common data sets for training ASR and TTS systems. The use of ASR data to train TTS systems is one of the most significant achievements of the project, and is described under WP2 below.

***Contribution to D2.1***

The common datasets identified in WP1 were crucial to the work conducted in WP2.

***Deliverable D1.2***

A list of considered languages, and the considerations made in their selection, was provided in this deliverable. A list of related resources (speech and text collections for each language considered) was also provided. These speech and text collections were used for the purposes of WP3.

***Deliverable D1.3***

This deliverable described various machine translation (MT) systems used to automatically translate datasets used in WP3. The advantages and disadvantages of each system were detailed. For example, some MT systems are appropriate for real-time demonstration, while others are more suitable for research.

***Year 3***

UCAM developed web-based syntactic hierarchical phrase-based SMT systems for integration into the EMIME demos. Given the constraints of the demonstration systems, it was judged that specialized translation engines would be more suitable than general purpose translation services such as provided by Google. Effort was devoted to developing translation grammars and language models capable of providing interactive translation services. A constrained translation grammar was developed based on the phrase inventory provided by Nokia. An MT API was providing for accessing these systems, which continue to run at UCAM.

***Data collection***

The data we have collected, and which is being publicly released, is described in Section 4.2 – B.2.

#### 4.1.3.2 WP2: Speaker adaptation and independence for speech recognition and synthesis

This workpackage was where much of the work in Year 1 took place. The core structure of EMIME comprised the development of speaker adaptation techniques in WP2 and their application to speech synthesis and speech recognition in a unified way, which lead into WP3 where these techniques were taken into a cross-lingual setting.

##### 4.1.3.2.1 Overall objectives of WP2

- The development of effective unsupervised speaker adaptation techniques that rely only on user input to the speech recognition system for adaptation data. Adaptation not only resulted in improvements to recognition performance, but also provided systematic improvement to speech synthesis, by adapting the models such that the synthesised speech bears the same identity characteristics as that of the input speaker. This required the modification and extension of existing speaker adaptation schemes for recognition to meet the additional requirements imposed by synthesis.

  *Progress in Year 1: our experiments showed that unsupervised adaptation works just as well as supervised adaptation for speech synthesis (see under "Unsupervised adaptation for TTS" below).*

  *Progress in Year 2: our work mostly looked at extending and evaluating the unsupervised adaptation of TTS and unified modelling for ASR and TTS (see "Unsupervised speaker adaptation and Unified models for speech recognition and synthesis" below.).*

- The investigation and development of novel speech representations that are specifically designed for a joint recognition and synthesis modelling paradigm. This was intended to result in improved speech synthesis quality, maintaining the identity of the speaker, while not degrading (and possibly improving) speech recognition performance. The proposed speech representations were intended to operate within the speaker adaptation framework developed as part of the first WP2 objective.

  *Progress in Year 1: exploration of a wide range of spectral parameterisations and dimensionalities for both ASR and TTS (see under "Comprehensive baseline results for ASR" and "Comprehensive baseline results for TTS" below).*

  *Progress in Year 2: We investigated feature domain transformations to provide ability to rapidly adapt our models to new speakers and/or improve synthesis quality (See "Data driven feature transforms below").*

- Research conducted in this workpackage was focused on achieving its goals *within a single language* and was only conducted for the first 18 months of the project. Research in WP3 was then responsible for extending the achievements of WP2 to the multilingual domain.

  *Progress in Year 1: The monolingual work was almost completed and the transition into WP3 had begun.*

  *Progress in Year 2: We had already made significant progress in applying much of the work conducted in WP2 to cross-lingual scenarios.*

#### 4.1.3.2.2 Significant results obtained in WP2

***Comprehensive baseline results for ASR***

In D2.1 we reported a substantial series of experiments in all four languages to compare many different system configurations. Conclusions were drawn regarding system design parameters such as the acoustic feature type and dimensionality, model topology, adaptation method, and so on. The goal of these experiments was to discover how large the gap between ASR and TTS is and to identify how the models used for each can be made closer, to facilitate the transfer of adaptation parameters from ASR models to TTS models.

***Comprehensive baseline results for TTS***

Synthesis systems were built for all four languages. Substantial listening tests were conducted for English and Mandarin, leading to a number of statistically significant conclusions (including the two key results: use of ASR data and supervised vs. unsupervised adaptation). A smaller test was conducted to validate the Finnish system.

***Training TTS models on ASR data***

An analysis of the phonetic coverage of ASR corpora was made, which revealed that, across many speakers, wide coverage (measured in terms of number of unique triphones seen) can be obtained. Although this is not as extensive coverage as in dedicated speech synthesis corpora, it is sufficient for training "average voice" models. In a large listening test for English (see D2.1 for full results), we found that a system trained on the SI84 set from WSJ0 significantly outperformed a system trained on a collection of high-quality synthesis voices, in terms of speaker similarity, naturalness and intelligibility. This is a very important result.

***Unsupervised adaptation for TTS***

As mentioned earlier, unsupervised adaptation can provide equivalent quality synthesis to supervised adaptation. This is also a very important result.

***Bridging the gap between ASR and TTS***

The extensive experiments mentioned above included configurations that were either oriented towards ASR, or towards TTS, or a compromise between the two. Whilst a number of system design parameters have very different optimal values for ASR compared to TTS, several aspects of the systems can be configured identically, including the pronunciation dictionary, phone set and parameter tying scheme.

***Deliverable D2.1***

The most substantial deliverable so far, this laid the foundation for future work in both WP2 and WP3.

***Unsupervised speaker adaptation and Unified models for speech recognition and synthesis***

We formally evaluated both two-pass and decision tree marginalisation approaches in the context of unsupervised adaptation for TTS. Our results showed that these methods allow us to adapt TTS models using ASR transcriptions without significant degradation to TTS. By contrast, ASR performance was degraded. We then started applying these methods to cross-lingual adaptation in WP3.

We also used the outcomes from our studies on unsupervised speaker adaptation in the first year in our 2009 Blizzard Challenge entry. This system was the first such entry that used a system developed in a completely unsupervised fashion. While not being judged to be the best system, the EMIME entry nonetheless performed well in comparison to most other entries and was a positive demonstration of the techniques developed thus far in the EMIME project.

***Speech features for joint modelling for recognition and synthesis: data driven feature transformations***

We conducted two novel studies concerning derivation of optimal features from data. In one, we reported on the derivation of optimal warping for fundamental frequency. We showed that a generalised logarithmic compression (moving towards to linear) proved to be preferable to the usual logarithmic scaling. This is consistent with our understanding of human perception in which the relationship between perceived and actual frequency is believed to be approximately linear at low frequencies.

In other work, we investigated the application of vocal tract length normalisation (VTLN) to speech synthesis. VTLN is applied using a bilinear transform, which is already incorporated into the mel-generalised analysis, hence, in this case VTLN may be considered equivalent to adaptation of the spectral warping. Our results showed successful application of this approach, especially for a small amount of adaptation data. This work continued in the framework of cross-lingual studies in WP3.

***Deliverable D2.2***

This was the final workpackage report for WP2 and was another substantial report, detailing the final achievements of WP2 at the midpoint of the EMIME project. This deliverable provided details of the above achievements with respect to unsupervised speaker adaptation and features for joint recognition and synthesis. In addition to directly addressing these key objectives of WP2, we also reported our progress in related research, namely, the development of algorithms for minimum generation error training and adaptation, autoregressive HMM for TTS, Bayesian acoustic modelling approach, discriminative training for TTS and flat-start TTS model training.

#### 4.1.3.3 WP3: Language independence and adaptation for speech recognition and synthesis

Whilst WP2 concentrated on building baseline systems and performing comprehensive experiments exploring the configuration of both ASR and TTS systems with the goal of finding as much common ground as possible, WP3 concerned cross-lingual adaptation. The role of WP3 in the early stages of the project was mainly to ensure that the methods and tools being used in the other workpackages are appropriate for the multilingual and cross-lingual work that took place later in WP3.

##### 4.1.3.3.1 Overall objectives of WP3

- The overall objective of WP3 was the further development of the monolingual unsupervised speaker adaptation techniques of WP2 in a cross-lingual setting. This produced effective cross-lingual speaker adaptation that modifies acoustic models using the input speech of a source language speaker and transforms the acoustic models used for target language synthesis in that speaker's voice. This was primarily achieved using cross-lingual acoustic adaptation techniques to provide cross-language linking of the recognition and synthesis systems developed in WP2. Speaker parameters can be learned either in supervised mode (making use

of reference speech transcriptions) or using unsupervised adaptation (making use of automatically-generated transcriptions).

*Progress in Year 1: Monolingual baseline systems were defined and implemented, based on the results from WP2 reported in D2.1. Identification of language-specific characteristics of the baseline systems. Identification of main differences between ASR and TTS models. Some initial cross-language experiments were carried out.*

*Progress in Year 2: Based on the intra-lingual results of WP2, the first supervised and unsupervised cross-lingual adaptation methods were proposed and implemented and initial experiments were performed.*

*Progress in Year 3: Advances in recognition, translation and synthesis, summarised under "Cross-lingual speaker adaptation" below*

- Multi-lingual acoustic HMMs will be developed for recognition and HTS in a cross-lingual setting so that speaker adaptation parameters learned in one language will be immediately available for personalised speech synthesis in a second language. The challenge will be to develop modelling techniques which ensure that the quality of these systems is comparable to monolingual synthesis. Multilingual recognition systems will also be developed using the multilingual acoustic HMMs and the recognition performance of these systems will be evaluated. Note that even though these multilingual recognition systems can be expected to lag in quality relative to monolingual recognition, they will still provide a means to perform unsupervised speaker adaptation by providing the parameters needed for personalised speech synthesis.

  *Progress in Year 2: Multilingual recognition systems have been implemented and tested in parallel with a new data mapping approach that where the target language recognition system can be directly applied for the adaptation data in the source language.*

  *Progress in Year 3: Advances in recognition, translation and synthesis, summarised under "Cross-lingual speaker adaptation" below*

#### 4.1.3.3.2 Significant results obtained in WP3

***D3.1***

The comparison and analysis of the HTK-based baseline ASR systems selected from WP2, working in four very different languages: English, Finnish, Mandarin, and Japanese.

***Cross-lingual speaker adaptation***

Several new methods for cross-lingual speaker adaptation in ASR and TTS were developed, tested and reported. The division of the cross-lingual adaptation methods to transform mapping, data mapping and multilingual ASR methods was proposed. The intra-lingual vs. cross-lingual and supervised vs. unsupervised adaptation schemes were compared and the developed state-mapping methods were described. Case studies were performed for mapping between various language pairs that are relevant to the project. A summary of our achievements is:

- Speech recognition
  - ASR of speech with foreign accent using cascaded transforms
  - Cross-lingual adaptation of ASR models to deal with foreign-accented speech
  - Mixed-language ASR
- Machine translation
  - Integration of machine translation and speech synthesis

  - Morphological analysis for Finnish and other morphologically-rich languages

- Speech synthesis adaptation

  - Real-time demonstrator now incorporates cross-lingual speaker adaptation – see WP4
  - VTLN for speech synthesis adaptation
  - Two-pass decision tree method for CLSA refined and published as a journal paper
  - Further analysis of the state mapping-based approach to CLSA
  - Voice conversion
  - Further analysis of the role of the average voice model in relation to the quality and attractiveness of adapted voices

- Advances in core speech synthesis techniques

  - We have made many contributions to statistical parametric speech synthesis, detailed in D3.3

#### 4.1.3.4 WP4: Integration and demonstration

This workpackage provided two demonstrator architectures for the project. The research demonstrator was created to facilitate research by all partners. It enables the construction of complete run-time systems using various combinations of available modules. These systems were evaluated as a whole and in parts, by WP5. The real-time demonstrator was created for use in a task-based evaluation by WP5 of cross-language speaker adaptation.

##### 4.1.3.4.1 Overall objectives of WP4

- Two types of demonstrators were designed and created in the project: a research demonstrator for formal evaluation of components, and a real-time embedded demonstrator for real-life user evaluations. The first objective of WP4 was to provide the necessary infrastructure (system architecture definition, interface definition, user interface) for these demonstrators.

  *Progress in Year 1: Research demonstrator architecture and interfaces defined. Detailed design of real time demonstrator underway (SIP protocol for IP connection, client-server architecture, a prototype user interface has been also proposed for discussion).*

  *Progress in Year 2: The research demonstrator was developed to improve the feasibility and practicality of research demonstration. Work in year 2 included performance improvements and enriched functionality, including machine translation and cross-lingual ASR-TTS acoustic model adaptation. It became easier to configure, integrate and test different components and models. The infrastructure of the first version of the real-time demonstrator was built, using a client-server architecture. The client software records and sends speech to the server to transcribe, translate, synthesise and return to the other client. The client UI also supports viewing and correcting misrecognized speech, sending the human-corrected transcription to machine translation directly to improve the system performance. The core tasks of ASR, MT and TTS are executed on a server, connected via GPRS, 3G or WiFi connection).*

  *Progress in Year 3: further optimisation of the infrastructure for the real-time demonstrator by providing capability for multiple pairs of users and also supporting a single-user use case.*

- The second objective of this workpackage was to carry out the actual development, integration and implementation work for the demonstrators.

  *Progress in Year 1: Research demonstrator implemented, tested by other partners, and upgraded according to the feedback received. The research demonstrator is a flexible tool written in a scripting language which*

*allows the construction of end-to-end systems from any combination of appropriate modules. It runs on any platform and supports both 32-bit and 64-bit machines (since both are in use within the consortium). Initial prototype of real-time demonstrator was made.*

*Progress in Year 2: In EMIME research demonstrator, all modules had been integrated together and shown to run smoothly. All necessary acoustic and language models were available for EMIME defined languages (English, Chinese, Japanese and Finnish).*

*Progress in Year 3: progressed from monolingual voice cloning to cross-lingual voice adaptation with speech to speech translation. All EMIME selected languages, English, Chinese, Japanese and Finnish, were now supported. The real-time demonstrator was extended to support multiple use cases and multi-threading. We also implemented a phrasebook version of the real-time demonstrator, in which cross lingual speaker adaptation has a fast implementation.*

#### 4.1.3.4.2 Significant results obtained in WP4

***Research demonstrator***

This provides a single architecture which all partners used to conduct evaluations (note: this does not include training statistical models – other tools have been developed in WP2 for this). The research demonstrator has been tested and verified by all partners. A "Voice-cloning Demo" system based on the research demonstrator was exhibited at the Interspeech 2009 conference and has also been used for public engagement. For cross-lingual speaker-adaptive speech-to-speech translation, the required acoustic and language models are available for all the desired languages (English, Chinese, Japanese and Finnish). It is an ideal platform for performance evaluations, in which each component can be evaluated individually or together within a complete end-to-end system. A multi-lingual version of the Voice Cloning demo is also implemented in this framework.

***Real-time demonstrator***

The EMIME real-time demonstrator makes use of GPRS, 3G or WiFi to connect and communicate between two clients in a similar way to VoIP (voice over IP). At the client end, the audio engine is used to play back synthesized speech. The audio send and receive modules are in charge of audio, control command and text transmission. The control module processes the interaction between client and server. At the server end, the control module processes the interaction between client and server and maintains the session between two mobile clients. The audio send and receive modules are in charge of audio, command and text transmission same as the client end. The core engine is used to execute the major tasks of the speech-to-speech translation, including cross-lingual and unified speaker adaptation. The UI component was designed and implemented as part of the client software.

### 4.1.3.5 WP5: Evaluation and dissemination

#### 4.1.3.5.1 Overall objectives of WP5

- External evaluation of system components.

  *Progress in Year 1: Organisation of the Blizzard Challenges in 2008 and 2009. Pre-EMIME baseline entered to Blizzard Challenge 2008. EMIME system entered to Blizzard Challenge 2009.*

  *Progress in Year 2: Organisation of the Blizzard Challenges in 2009 and 2010. EMIME system entered to Blizzard Challenge 2009 (details in paper published at the Blizzard 2009 workshop). EMIME system entered to Blizzard Challenge 2010.*

  *Progress in Year 3: Organisation of the Blizzard Challenges in 2010, 2011 and 2012. EMIME systems entered to Blizzard Challenge 2010 (details in paper published at the Blizzard 2010 workshop) and Blizzard Challenge 2011 (underway at the time of writing).*

- Internal evaluation of system components, and research prototypes of the full system.

  *Progress in Year 1: Definition of evaluation methods for ASR and TTS. Large-scale listening tests of the baseline speech synthesis systems for English and Mandarin, plus a small-scale test for Finnish. Internal testing of the research demonstrator architecture had begun.*

  *Progress in Year 2: extensive component-level and end-to-end evaluations reported in D2.2 and D3.2.*

  *Progress in Year 3: extensive component-level and end-to-end evaluations reported in D3.3 and D5.5.*

- Internal evaluation of the real-time demonstrator.

  *Progress in Year 2: real-time demonstrator testing; the first stage was to replicate the demonstrator at UEDIN and AALTO sites.*

  *Progress in Year 3: demonstrator fully functional and used in a small-scale usability study.*

#### 4.1.3.5.2 Significant results obtained in WP5

***Website***

***Project website***

The project website was a crucial tool for inter-project communication and collaboration. It contains copies of all important documents, an ongoing record of the project meetings and telephone calls and areas for each workpackage to record progress and plan future work. The website also features material of interest to a general audience, including demonstrations of speech synthesis in several languages, including interactive (text-to-speech) demonstrations for some languages running on a server at UEDIN. The project has released foreground as open source software and data: this is available on the website, along with all our publications.

***Collaborative authoring tools***

UEDIN provided a Subversion-based repository for Latex documents that allows simultaneous authoring and version control. This system was used to produce all deliverables and many of the co-authored publications.

***Definition of evaluation criteria***

For ASR, the standard metric of WER was adopted. For TTS, both objective measures (mel-cepstral distance, RMSE of F0 and voiced/unvoiced error) and subjective measures (based on the intelligibility, naturalness and speaker similarity tests used in the Blizzard Challenge) were adopted.

***Listening tests for WP2***

The tests for English and Mandarin were large (around 100 listeners each – comparable to the paid-listener subsets used in the full Blizzard listening test) and provided many useful results. The large listener pool meant that we were able to find statistically significant differences between many systems.

***Blizzard Challenge 2009***

Our entries performed well.

***Blizzard Challenge 2010***

The languages used were relevant to EMIME – English and Mandarin – with high-quality databases being contributed by the Chinese Academy of Sciences, Phonetic Arts and UEDIN. Our entries performed well.

***Blizzard Challenge 2011 and 2012***

Both the 2011 and 2012 challenges are underway at the time of writing. The 2011 challenge is relatively straightforward (one language, large high quality database) and of only limited relevance to EMIME. The 2012 challenge involves the use of audiobook data, where the techniques for dealing with noisy, or otherwise more than usually challenging, data that we have developed in EMIME will be relevant.

***Overall evaluations for WP3 and WP4***

The evaluation of the research demonstrator was the main focus, since comprehensive laboratory-based results can be obtained for the final versions of our methods. The evaluation of the realtime demonstrator was less formal and extensive, partly because the reviewers in Year 2 recommended not to focus heavily in this demonstrator.

The evaluation of the research demonstrator takes the form of a carefully designed sequence of four experiments, starting with listeners' evaluations of speaker identity for human speech and concluding with evaluations of the cross-lingual speaker-adapted synthetic speech. A final, fifth, experiment, deals with within-language speaker discrimination rather than across-language speaker discrimination.

In going from Exp. I to Exp. IV an increasing number of the elements that play a role in the EMIME scenario are included. Exp. IV should be viewed as the final EMIME evaluation: listeners were asked to decide on a speaker's identity whilst comparing a user's Mandarin natural speech to their synthetic English speech – which had been adapted to the user's voice by using their Mandarin natural speech as adaptation data. This is a very challenging task for listeners as they have to deal with the combination of across-language (Mandarin versus English) and across speech type (natural versus synthetic) factors while trying to identify speakers. We carried out the intermediate experiments to answer questions about how listeners deal with these various factors.

Exp. I investigates how well listeners discriminate between bilingual speakers across languages. Exp. II focuses on how well listeners are able to identify speakers when the trials consist of synthetic speech instead of natural speech. In Exp. III, we compare natural Mandarin to synthetic English – created using within-language adaptation (the sentences used for adaptation are from the same language as the synthetic speech that is being created). In Exp. IV, the comparison in still between natural Mandarin and synthetic English, but in this case, across-language adaptation has been applied (sentences from Mandarin are used to adapt English synthetic speech). These experiments touch on the question: Does the synthetic speech which has been adapted to sound like the original speaker actually sound like them? As Exp. II – IV all deal with across-language speaker discrimination, the results cannot separate out whether changes in the listeners' performance are due to across-language issues or across-speech type factors. Exp. V fills this gap by only looking at speaker discrimination across-speech types. It has a slightly different focus as it deals with within-language speaker discrimination and additionally the effect of accent is investigated.

The real-time demonstrator described in D4.6 was evaluated internally using a task-based evaluation. A relatively small-scale test has been conducted in order to demonstrate that we have achieved the goal of a functioning real-time system which incorporates cross-lingual speaker adaptation. Three pairs of users were given a simple scenario which they were asked to complete using their own languages. Subjects found the system reasonably easy to use. They were able to understand the synthetic speech. Although they did not think the translated speech sounded like themselves, they did think that the other person's translated speech sounded a bit more like them. Self-perception of speaker identity is something we considered investigating, but ruled out because of the logistical difficulty of arranging an experiment where subjects are both talkers and, later, listeners. Most of the subjects were satisfied with the system's speed and they found the phone easy to navigate.

***Dissemination***

Demonstrations of EMIME systems were given at ACL 2010 and SSW7 (the seventh ISCA speech synthesis workshop). A meeting involving academic and commercial groups (including Toshiba and Phonetic Arts) was held in Cambridge, UK just after the Year 2 review meeting. Many presentations have been given about EMIME, listed in this document, and we have received plenty of media attention.

***Awards***

Junichi Yamagishi (UEDIN) was awarded the 2010 Itakura Prize for Innovative Young Researchers by the Acoustical Society of Japan for his contributions to "Speaker adaptation techniques for speech synthesis".

### 4.1.4 Potential impact

EMIME made substantial advances in adaptive statistical parametric speech synthesis, and has developed useful applications based on new techniques, including our primary target of cross-lingual speaker adaptation. Whilst the principal aim of the project was to create personalised speech-to-speech translation, the technology we have created has a much wider range of application and could lead in a number of directions. Details of the following results can be found in the scientific publications, listed on the project website, and implementations are available in the open source software we have released.

#### Core speech synthesis

***Results*** Robust techniques for using lower-quality data, such as ASR corpora or recordings made in domestic environments. Methods for building voices on high sample rate data, using alternative auditory scales and higher-order spectral representations. Contributions to minimum generation error (MGE) training. Flat-start training. Auto-Regressive HMM, which is an alternative to the Trajectory HMM. Parameter clustering methods for AR-HMMs. Bayesian approach to HMM synthesis.

***Use*** These results will be useful to any speech synthesis company using statistical parametric methods (which most now are).

***Commercial impact*** There is the potential for substantial impact through general improvements to the quality of speech synthesis.

***Societal impact*** Impacts more widely will be as a consequence of improved quality speech synthesis, delivered as commercial products and services.

#### Cross-lingual speaker-adaptive speech synthesis

***Results*** Unsupervised adaptation (using untranscribed speech data). Two-pass decision tree method and decision tree marginalisation methods for cross-lingual speaker adaptation. State mapping-based approach to cross-lingual speaker adaptation. Further analysis of the role of the average voice model in relation to the quality and attractiveness of adapted voices. Some work on voice conversion. Cross-language speaker adaptation has been demonstrated for several pairs of languages.

***Use*** Techniques for speaker adaptation can in general be also applied to other forms of adaptation, such as to speaking styles or expressive speech generation. The primary use case is for building voices from "imperfect" data: small quantities, possibly noisy, etc.

***Commercial impact*** Widespread and immediate possibilities exist for adaptive speech synthesis and we are likely so see the takeup of our methods by speech synthesis companies in the short term. The prospect of making many different voices without the cost of current speaker-dependent voices (which is estimated to be in the range €10–100k per voice) is highly attractive to speech synthesis companies and their customers.

***Societal impact*** Making better speech synthesis available for more languages. Personalised speech-to-speech translation. Information access. Communication across languages.

### Rapid adaptation for speaker-adaptive speech synthesis

***Results*** VTLN for speech synthesis adaptation

***Use*** Rapid adaptation with as little as one sentence. Low computational complexity.

***Commercial impact*** Personalised speech synthesis embedded into real applications where very little speech data are available and very rapid adaptation is desired

### Resources created to assist exploitation

***Results*** A number of trained models for the EMIME languages will be released for others to use. High-quality speaker-dependent and speaker-adaptive systems have been demonstrated for several languages. Many improvements and bug fixes to the HTS and Festival toolkits. Many of our methods are now included in the public release version of the HTS toolkit, or will be in future releases. The project website has a "Downloads" section, where many useful tools and databases can be freely downloaded. All our publications are also listed on the website, with links to no-cost public access versions wherever possible.

***Use*** Our methods for cross-lingual speaker adaptation are published in academic papers and released as software, enabling their use and exploitation by others. The primary application of our techniques is in speech-to-speech translation, although other uses could be found.

***Commercial impact*** Use of some models is restricted according to the license for the training data, although commercial licenses can be obtained for all databases, so commercial use is possible in principle. Software is open source and available for commercial exploitation. Some restrictions apply to certain components, such as the vocoder and some lexicons, but commercial licences are available for these, and/or other commercially available alternatives can be substituted.

### Application to speech-to-speech translation

***Results*** This was the primary goal of EMIME and we have demonstrated that it is possible to personalise the spoken output of a speech-to-speech translation device.

***Use*** Speech-to-speech translation is a widely-researched topic and our method could 'plug in' to any existing system which uses HMM-based speech recognition.

***Commercial impact*** We are starting to see commercial speech-to-speech translation applications, particularly for mobile devices (as targeted by EMIME). Personalised versions of such systems are now possible.

***Societal impact*** The societal impacts of speech-to-speech translation can only be increased by the use of personalisation.

### Assistive technology applications

***Results*** Our techniques for building voices from small amounts of noisy data from non-professional speakers, yet still achieving high quality output, will enable the next generation of voice-output communication aids.

***Use*** Assistive devices for people who are losing their voice, in which the device speaks with their original voice. Also, devices for people who cannot speak, with a very large 'library' of voices to choose from in terms of gender, accent, age, social class, etc.

***Commercial impact*** There is a niche market for voice-output assistive technology, with key players including Tobii (Sweden), Toby Churchill (UK) and Dynavox (USA). UEDIN are in discussions with Tobii regarding the use of speaker-adaptive synthesis on their devices and an initial proof-of concept has been produced for a patient with Motor Neurone Disease. This research is continuing and is funded.

***Societal impact*** The loss of the voice has a profound affect not only on a person's ability to speak, but also their identity. Using impersonal voice-output communication aids is often unattractive, because of the very limited range of voices available. Personalised communication aids are very often requested by patients. Technology created in EMIME can deliver this.

### Automatic speech recognition

***Results*** ASR of speech with foreign accent using cascaded transforms. Cross-lingual adaptation of ASR models to deal with foreign-accented speech. Mixed-language ASR. Deterministic annealing based training algorithm for Bayesian speech recognition

***Use*** Whenever automatic speech recognition must deal with non-native speakers or operate in a mixed-language environment, both of which are common in the European context.

***Commercial impact*** Improvements to ASR have obvious commercial application.

***Societal impact*** Spoken interaction systems will be better able to handle accented speech.

### Advances in machine translation

***Results*** Integration of machine translation and speech synthesis. Morphological analysis for Finnish and other morphologically-rich languages

#### 4.1.4.0.3 Main dissemination activities and exploitation of results

The primary forms of dissemination we chose were 1) publication in academic journals and conference proceedings (which can all be freely accessed via the project website) and 2) the release of open source software and data which allows others to replicate our results and exploit them commercially. We considered that it would be more effective to make a public release of this intellectual property, rather than seek exclusive licensing opportunities with individual companies. Making our methods available in the form of software implementations allows others to more rapidly adopt and improve upon them. It also enables continued academic research largely free from usage restrictions on the tools and data. Of course, some of the techniques we created have built upon the work of others, and we have used some external data. These small parts of our system can be obtained by others from the original sources, or can be replaced with suitable alternatives. In addition to publications, software, data and the project website, we have also given a large number of presentations about the project to both academic and commercial audiences, and have presented our work to the wider public through media exposure and some engagement activities with schools.

Section 4.1.4 suggests how each of the technological and scientific outputs of the project could be exploited and how that could benefit society. We are also continuing to work ourselves on further developments of this technology and are finding new uses for the methods we invented. One notable use is in assistive technology, where our methods are able to make personalised voices for alternative and augmentative communication (AAC) devices.

One particularly exciting breakthrough we have made since the EMIME project is that the statistical modelling method can be used to reconstruct disordered voices, thus creating intelligible synthetic speech even when the only recordings we can obtain of a patient are of disordered (and hard to understand) speech. One limitation we found in the EMIME project is that it is relatively difficult and expensive to tackle a new language in speech synthesis, because of the linguistic resources required (e.g., text processor, labelled data). Some members of the consortium plan to attack this problem next, by using machine learning and recent advances from unsupervised NLP, working from unlabelled speech and text data.

### 4.1.5 The address of the project public website and project logo

www.emime.org


#### 4.1.5.1 Contact details

| Organisation | Contact | | Country |
|---|---|---|---|
| University of Edinburgh | Simon King | Simon.King@ed.ac.uk | UK |
| Idiap Research Institute | Phil Garner | Phil.Garner@idiap.ch | Switzerland |
| Aalto University | Mikko Kurimo | Mikko.Kurimo@tkk.fi | Finland |
| Nagoya Institute of Technology | Keiichi Tokuda | tokuda@nitech.ac.jp | Japan |
| Nokia Corporation | Jilei Tian | jilei.tian@nokia.com | Finland |
| University of Cambridge | Bill Byrne | bill.byrne@eng.cam.ac.uk | UK |

# 4.2

# Use and dissemination of foreground

A plan for use and dissemination of foreground (including socio-economic impact and target groups for the results of the research) shall be established at the end of the project. It should, where appropriate, be an update of the initial plan in Annex I for use and dissemination of foreground and be consistent with the report on societal implications on the use and dissemination of foreground (section 4.3 H). The plan should consist of:

## 4.2 – A Dissemination measures

Links to openly accessible versions of publications, with permanent identifiers, can be found via `http://www.emime.org/learn/publications`.

### 4.2 – A.1 List of all scientific (peer reviewed) publications relating to the foreground of the project

The following conference papers have arisen from work conducted wholly or partially within EMIME, including experiments conducted by members of the consortium as precursors to the work in EMIME:

- Junichi Yamagishi, Takashi Nose, Heiga Zen, Tomoki Toda, Keiichi Tokuda, Takao Kobayashi, "Performance evaluation of average-voice based speech synthesis system for Blizzard Challenge 2007," Spring Meeting of ASJ, vol.I, 2-11-3, pp.339–342, March 2008.
- Heiga Zen, Keiichi Tokuda, "An HMM-based speech synthesis system," The 70th National Convention of IPSJ, vol.5, 4L-6, pp.5-359–5-360, March 2008.
- Yi-Jian Wu, Heiga Zen, Yoshihiko Nankaku, Keiichi Tokuda, "Minimum generation error criterion considering global/local variance for HMM-based speech synthesis," 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), pp.4621–4624, Las Vegas, Nevada, U.S.A., March 30-April 4, 2008.
- Junichi Yamagishi, Takashi Nose, Heiga Zen, Tomoki Toda, Keiichi Tokuda, "Performance evaluation of the speaker-independent HMM-based speech synthesis system HTS-2007 for the Blizzard Challenge 2007," 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), pp.3957–3960, Las Vegas, Nevada, U.S.A., March 30-April 4, 2008.
- J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda and K. Tokuda, The HTS-2008 system: yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. Blizzard Challenge 2008 Workshop, Brisbane, Australia, Sept. 2008
- Y.-J. Wu and K. Tokuda. Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis, Proc. of Interspeech 2008, pp. 577-580, Brisbane, Australia, Sept. 2008

- J. Yamagishi, Z. Ling, and S. King. Robustness of HMM-based Speech Synthesis. Proc. Interspeech 2008, pp. 581-584, Brisbane, Australia, Sept. 2008
- Simon King, Keiichi Tokuda, Heiga Zen, and Junichi Yamagishi. Unsupervised adaptation for HMM-based speech synthesis. pages 1869-1872, Brisbane, Australia, September 2008.
- Yi-Jian Wu, Keiichi Tokuda, "HMM training by minimizing log spectral distortion between generated and original LSPs for speech synthesis," Autumn Meeting of ASJ, vol.I, 1-4-6, pp.249–250, September 2008.
- Kei Hashimoto, Heiga Zen, Yoshihiko Nankaku, Keiichi Tokuda, "HMM based speech synthesis using cross validation for Bayesian criterion," Autumn Meeting of ASJ, vol.I, 1-4-7, pp.251–252, September 2008.
- Keiichiro Oura, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda, "Tying variance for HMM-based speech synthesis," Autumn Meeting of ASJ, vol.I, 2-p-29, pp.421–422, September 2008.
- Tomoki Toda, Keiichi Tokuda, "Probabilistic speech spectral analysis based on factor analyzed trajectory HMM," Autumn Meeting of ASJ, vol.I, 3-4-6, pp.293–294, September 2008.
- Oliver Watts, Junichi Yamagishi, Kay Berkling, and Simon King. HMM-based synthesis of child speech. In Proc. of The 1st Workshop on Child, Computer and Interaction (ICMI'08 post-conference workshop), Crete, Greece, October 2008
- Zhi-Peng Yu, Yi-Jian Wu, Heiga Zen, Yoshihiko Nankaku, Keiichi Tokuda, "Analysis of stream-dependent tying structure for HMM-based speech synthesis," International Conference on Signal Processing (ICSP'08), pp.655–658, Beijing, China, October 26-29, 2008.
- Keiichiro Oura, Yoshihiko Nankaku, Tomoki Toda, Keiichi Tokuda, Rannierry Maia, Shinsuke Sakai, Satoshi Nakamura, "Simultaneous Acoustic, Prosodic, and Phrasing Model Training for TTS Conversion Systems," International Symposium on Chinese Spoken Language Processing (ISCSLP2008), SPE1.1, pp.1–4, Kunming, China, December 16-19, 2008 (Best Student Paper Award).
- Yi-Jian Wu, Simon King and Keiichi Tokuda, "Cross-Lingual Speaker Adaptation for HMM-based Speech Synthesis," International Symposium on Chinese Spoken Language Processing (ISCSLP2008), SPE1.1, pp.9–12, Kunming, China, December 16-19, 2008.
- Kei Hashimoto, Heiga Zen, Yoshihiko Nankaku, Keiichi Tokuda, "Bayesian Context Clustering Using Cross Validation for HMM-Based Speech Synthesis," IEICE Technical Report, 2008-SLP-74-13, December 2008.
- Keiichiro Oura, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda, "Tying covariance parameters for HMM-based speech synthesis," IEICE Technical Report, 2008-SLP-74-37, December 2008.
- Akinobu Lee and Tatsuya Kawahara, "Developing a speech recognition interface using Julius," vol.11, no.1, pp.31–38, February 2009.
- Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda, "Hidden semi-Markov model based Bayesian speech synthesis," Spring Meeting of ASJ, vol.I, 1-6-7, pp.303–304, March 2009.
- Shinji Sako, Keiichi Tokuda, Tadashi Kitamura, "VoiceMaker: A simplified toolkit to build acoustic model for HMM speech synthesis," Spring Meeting of ASJ, vol.I, 1-6-8, pp.305–306, March 2009.
- Kyohei Nagao, Heiga Zen, Yoshihiko Nankaku, Keiichi Tokuda, "Investigation of Global Variance Modeling for HMM-Based Speech Synthesis," Spring Meeting of ASJ, vol.I, 1-R-19, pp.427–428, March 2009.
- Kei Hashimoto, Heiga Zen, Yoshihiko Nankaku, Takashi Masuko, and Keiichi Tokuda, "A Bayesian approach to HMM-based speech synthesis," Proc. of ICASSP 2009, Taipei, Taiwan, April 19-24, 2009.
- Yi-Jian Wu, Keiichi Tokuda "Minimum generation error training by using original spectrum as reference for log spectral distortion measure," 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), Taipei, Taiwan, April 19-24, 2009.

- Lu Heng, Wu Yi-Jian, Tokuda Keiichi, Dai Li-Rong, Wang Ren-Hua, "FULL covariance state duration modeling for HMM-based speech synthesis," 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), Taipei, Taiwan, April 19-24, 2009.
- J. Tian and J. Nurminen, "Optimization of text database using hierachical clustering," Proc. IEEE ICASSP 2009, Taipei, Taiwan, April 2009
- Mikko Kurimo, Teemu Hirsimäki, Ville Turunen, Sami Virpioja, and Niklas Raatikainen. "Unsupervised decomposition of words for speech recognition and retrieval, " in Proceedings of the 13th International Conference Speech and Computer, SPECOM 2009, pages 23-28, St. Petersburg, Russia, June 21-25 2009.
- A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. In Procedings of NAACL-HLT, 2009.
- Teemu Hirsimäki and Mikko Kurimo. Analysing Recognition Errors in Unlimited-Vocabulary Speech Recognition. In Procedings of NAACL-HLT, 2009.
- J. Yamagishi, M. Lincoln, S. King, J. Dines, M. Gibson, J. Tian, Y. Guan, "Analysis of Unsupervised and Noise-Robust Speaker-Adaptive HMM-Based Speech Synthesis Systems toward a Unified ASR and TTS Framework," Proc. Blizzard Challenge 2009 (Edinburgh, U.K.).
- Keiichiro Oura, Yi-Jian Wu, and Keiichi Tokuda, "Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2009," Proc. Blizzard Challenge 2009 (Edinburgh, U.K.).
- M. Gibson, "Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models,", Proc. of Interspeech 2009 (Brighton, U.K.).
- M. Shannon and W. Byrne, "Autoregressive HMMs for speech synthesis", Proc. of Interspeech 2009 (Brighton, U.K.).
- Keiichiro Oura, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda, "Tying covariance matrices to reduce the footprint of HMM-based speech synthesis systems," Proc. of Interspeech 2009 (Brighton, U.K.).
- Yi-Jian Wu, Long Qinz, Keiichi Tokuda, "An improved minimum generation error based model adaptation for HMM-based speech synthesis," Proc. of Interspeech 2009 (Brighton, U.K.).
- Yi-Jian Wu, Yoshihiko Nankaku, Keiichi Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," Proc. of Interspeech 2009 (Brighton, U.K.).
- Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda, "A Bayesian Approach to Hidden Semi-Markov Model Based Speech Synthesis," Proc. of Interspeech 2009 (Brighton, U.K.).
- Sayaka Shiota, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda, "Deterministic Annealing based Training Algorithm for Bayesian Speech Recognition," Proc. of Interspeech 2009 (Brighton, U.K.).
- Junichi Yamagishi, Bela Usabaev, Simon King, Oliver Watts, John Dines, Jilei Tian, Rile Hu, Yong Guan, Keiichiro Oura, Keiichi Tokuda, Reima Karhila, Mikko Kurimo, "Thousands of Voices for HMM-based Speech Synthesis," Proc. of Interspeech 2009 (Brighton, U.K.).
- John Dines, Junichi Yamagishi, Simon King, "Measuring the Gap Between HMM-based ASR and TTS," Proc. Interspeech 2009 (Brighton, U.K.).
- John Dines, Lakshmi Saheer, Hui Liang, "Speech recognition with speech synthesis models by marginalising over decision tree leaves," Proc. Interspeech 2009 (Brighton, U.K.).
- Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. "Overview and results of Morpho Challenge 2009," Working Notes for the CLEF 2009 Workshop, Corfu, Greece, September 2009.

- Yong Guan, Jilei Tian, "Evaluation of Flat Start Labeling for Phoneme based Mandarin HTS System," Proc. of OCOCOSDA2009.
- Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. "Overview and results of Morpho Challenge 2009," Multilingual Information Access Evaluation Vol. I, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 – October 2, 2009, Revised Selected Papers, Lecture Notes in Computer Science. Springer, 2010.
- Heiga Zen, Keiichiro Oura, Takashi Nose, Junichi Yamagishi, Shinji Sako, Tomoki Toda, Takashi Masuko, Alan W. Black, Keiichi Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," 2009 APSIPA Annual Summit and Conference, Sapporo Convention Center, Sapporo, Japan, October 5 - 7, 2009.
- Lakshmi Saheer, Philip N. Garner, John Dines, Hui Liang, "VTLN adaptation for statistical speech synthesis", Proc. ICASSP 2010 (Dallas, USA).
- Hui Liang, John Dines, Lakshmi Saheer, "A Comparison of Supervised and Unsupervised Cross-Lingual Speaker Adaptation Approaches for HMM-Based Speech Synthesis", Proc. ICASSP 2010 (Dallas, USA).
- M. Gibson, T. Hirsimäki, R. Karhila, M. Kurimo and W. Byrne, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction", Proc. ICASSP 2010 (Dallas, USA).
- Keiichiro Oura, Keiichi Tokuda, Junichi Yamagishi, Simon King, Mirjam Wester, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis," Proc. ICASSP 2010 (Dallas, USA).
- Kyosuke Kazumi, Yoshihiko Nankaku, Keiichi Tokuda, "Factor analyzed voice models for HMM-based speech synthesis," Proc. ICASSP 2010 (Dallas, USA).
- J. Yamagishi, S. King "Simple methods for improving speaker-similarity of HMM-based speech synthesis," Proc. ICASSP 2010 (Dallas, USA).
- Mikko Kurimo, William Byrne, John Dines, Philip N. Garner, Matthew Gibson, Yong Guan, Teemu Hirsimäki, Reima Karhila, Simon King, Hui Liang, Keiichiro Oura, Lakshmi Saheer, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, Mirjam Wester, Yi-Jian Wu, Junichi Yamagishi "Personalising speech-to-speech translation in the EMIME project," Proc. ACL 2010 System Demonstrations, Uppsala, Sweden, 13 July 2010.
- Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus, "Morpho challenge 2005-2010: Evaluations and results," Proc. 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology, Uppsala, Sweden, July 2010.
- Sami Virpioja, Jaakko Vyrynen, Andre Mansikkaniemi, and Mikko Kurimo, "Applying morphological decompositions to statistical machine translation," Proc. of the ACL 2010 5th Workshop on Statistical Machine Translation. ACL, July 2010.
- Kei Hashimoto, Yoshihiko Nankaku and Keiichi Tokuda, "Bayesian Speech Synthesis Framework Integrating Training and Synthesis Processes," Proc. SSW7, Kyoto, Japan, 2010.
- Shinji Takaki, Yoshihiko Nankaku and Keiichi Tokuda, "Spectral modeling with contextual additive structure for HMM-based speech synthesis," Proc. SSW7, Kyoto, Japan, 2010.
- Mirjam Wester, John Dines, Matthew Gibson, Hui Liang, Yi-Jian Wu, Lakshmi Saheer, Simon King, Keiichiro Oura, Philip N. Garner, William Byrne, Yong Guan, Teemu Hirsimäki, Reima Karhila, Mikko Kurimo, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda and Junichi Yamagishi, "Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project," Proc. SSW7, Kyoto, Japan, 2010.

- Lakshmi Saheer, John Dines, Philip N. Garner and Hui Liang, "Implementation of VTLN for Statistical Speech Synthesis," Proc. SSW7, Kyoto, Japan, 2010.
- Junichi Yamagishi and Oliver Watts, "The CSTR/EMIME HTS system for Blizzard Challenge 2010" Proc. Blizzard Challenge 2010 (Kyoto, Japan).
- Akira Saito, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda, "Voice activity detection based on conditional random fields using multiple features," Proc. Interspeech 2010 (Tokyo, Japan).
- M. Shannon and W. Byrne, "Autoregressive clustering for HMM speech synthesis," Proc. Interspeech 2010 (Tokyo, Japan).
- M. Wester, "Cross-lingual talker discrimination," Proc. Interspeech 2010 (Tokyo, Japan).
- Hui Liang and John Dines, "An Analysis of Language Mismatch in HMM State Mapping-Based Cross-Lingual Speaker Adaptation," Proc. Interspeech 2010 (Tokyo, Japan).
- Toyohiro Hayashi, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda, "Speaker Adaptation Based on Nonlinear Spectral Transform for Speech Recognition," Proc. Interspeech 2010 (Tokyo, Japan).
- Junichi Yamagishi, Oliver Watts, Simon King and Bela Usabaev, "Roles of the Average Voice in Speaker-adaptive HMM-based Speech Synthesis," Proc. Interspeech 2010 (Tokyo, Japan).
- Xianglin Peng, Keiichiro Oura, Yoshihiko Nankaku, Keiichi Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis considering differences between language-dependent average voices," Proc. ICSP 2010.
- Peter Smit and Mikko Kurimo, "Using stacked transformations for recognizing foreign accented speech," Proc. ICASSP 2011 (Prague, Czech Republic).
- Mirjam Wester and Reima Karhila, "Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation," Proc. ICASSP 2011 (Prague, Czech Republic).
- Kei Hashimoto, Junichi Yamagishi, William Byrne, Simon King, Keiichi Tokuda, " An analysis of machine translation and speech synthesis in speech-to-speech translation system," Proc. ICASSP 2011 (Prague, Czech Republic).
- Shinji Takaki, Keiichiro Oura, Yoshihiko Nankaku, Keiichi Tokuda, "An optimization algorithm of independent mean and variance parameter tying structures for HMM-based speech synthesis," Proc. ICASSP 2011 (Prague, Czech Republic).
- Shifeng Pan, Yoshihiko Nankaku, Keiichi Tokuda, Jianhua Tao, "Global variance modeling on frequency domain delta LSP for HMM-based speech synthesis," Proc. ICASSP 2011 (Prague, Czech Republic).
- Sandra Andraszewicz, Junichi Yamagishi, Simon King "Vocal Attractiveness Of Statistical Speech Synthesisers," Proc. ICASSP 2011 (Prague, Czech Republic).


The following technical reports have arisen from work conducted within EMIME:


- M. Shannon and W. Byrne, "A formulation of the autoregressive HMM for speech synthesis," Department of Engineering, University of Cambridge, UK, Technical Report CUED/F-INFENG/TR.629, 2009.
- M. Wester. "The EMIME Bilingual Database," Technical Report EDI-INF-RR-1388, The University of Edinburgh, 2010.
- M. Wester and H. Liang, "The EMIME Mandarin Bilingual Database," Technical Report EDI-INF-RR-1396, The University of Edinburgh, 2011.


The following journal papers have arisen from work conducted wholly or partially within EMIME:

- Akinobu Lee, "Large Vocabulary Continuous Speech Recognition Engine Julius," IEICE Information and System Society Journal, vol.13, no.4, February 2009.
- Teemu Hirsimäki, Janne Pylkkönen and Mikko Kurimo. "Importance of High-Order N-gram Models in Morph-Based Speech Recognition". IEEE Trans. Audio, Speech and Language Processing, Volume 17, Number 4, May 2009, pp. 724-732.
- Heiga Zen, Keiichi Tokuda, "TechWare: HMM-Based Speech Synthesis Resources," IEEE Signal Processing Magazine 26(4), pp 95–97, July 2009
- Junichi Yamagishi, Takashi Nose, Heiga Zen, Zhenhua Ling, Tomoki Toda, Keiichi Tokuda, Simon King, Steve Renals "A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis," IEEE Trans. Audio, Speech and Language Processing, Vol. 17, no. 6, pp 1208–1230, August 2009.
- Heiga Zen, Keiichi Tokuda, Alan W. Black, "Statistical parametric speech synthesis," Speech Communication 51(11), pp 1039-1064, November 2009.
- Keiichiro Oura, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee and Keiichi Tokuda, "A covariance-tying technique for HMM-based speech synthesis," IEICE Transactions on Information Systems, vol.E93-D, no.3, March 2010.
- Oliver Watts, Junichi Yamagishi, Simon King and Kay Berkling "Synthesis of Child Speech with HMM Adaptation and Voice Conversion," IEEE Trans. Audio, Speech and Language Processing, Vol. 18, Issue 5, pp 1005-1016, 2010.
- Junichi Yamagishi, Bela Usabaev, Simon King, Oliver Watts, John Dines, Jilei Tian, Rile Hu, Yong Guan, Keiichiro Oura, Keiichi Tokuda, Reima Karhila and Mikko Kurimo "Thousands of Voices for HMM-Based Speech Synthesis-Analysis and Application of TTS Systems Built on Various ASR Corpora" IEEE Trans. Audio, Speech and Language Processing, Vol. 18, Issue 5, pp 984-1004, 2010.
- John Dines, Junichi Yamagishi, Simon King, "Measuring the gap between HMM-based ASR and TTS,"IEEE Selected Topics in Signal Processing, Vol. 4, Issue 6, pp 1046-1058 December 2010.
- Matthew Gibson and William Byrne, "Unsupervised intralingual and cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," IEEE Transactions on Audio, Speech and Language Processing, Vol. 19, Issue 4, pp 895-904, 2011.
- Kei Hashimoto, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda, "Bayesian context clustering using cross validation for speech recognition," IEICE TRANSACTIONS on Information & Systems, vol.E94-D, no.3, pp.668–678, March 2011.
- Sayaka Shiota, Kei Hashimoto, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda, "Speech recognition based on statistical models including multiple phonetic decision trees" Acoustical Science and Technology (in press).
- A. Stan, J. Yamagishi, S. King and M. Aylett, "The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate," Speech Communication, vol.53, issue 3, pp.442–450, March 2011.


The following papers (authors and provisional titles are given) have been submitted to Interspeech 2011:

- Ling-Hui Chen, Yoshihiko Nankaku, Heiga Zen, Keiichi Tokuda, Zhen-Hua Ling, Li-Rong Dai "Estimation of Window Coefficients for Dynamic Feature Extraction for HMM-based Speech Synthesis"
- Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda, "Multi-Speaker Modeling with Shared Prior Distributions and Model Structures for Bayesian Speech Synthesis"
- Takafumi Hattori, Lei Li, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda "Speaker Recognition Based on GMMs Using Multiple Model Structures"

- Lei Li, Yoshihiko Nankaku, Keiichi Tokuda “A Bayesian Approach to Voice Conversion Based on GMMs Using Multiple Model Structures”
- Sayaka Shiota, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda “Acoustic modeling based on multiple model structures for Bayesian speech recognition”
- Ulpu Remes, Yoshihiko Nankaku, Keiichi Tokuda “GMM-based missing-feature reconstruction on multi-frame windows”
- R. Karhila and D.R. Sanand “Investigating the use of VTLN in Speaker Adaptation for Finnish Statistical Speech Synthesis”
- Janne Pylkkönen “Robustness of Discriminatively Trained Acoustic Models in Large Vocabulary Continuous Speech Recognition”
- Janne Pylkkönen “Analysis of Extended Baum-Welch and Constrained Optimization for Discriminative Training of HMMs”
- R. Karhila and M. Wester “Rapid Adaptation of Foreign-accented HMM-based Speech Synthesis”
- D. R. Sanand and M. Kurimo “A Study on Combining VTLN and SAT to Improve the Performance of Automatic Speech Recognition”
- H. Liang and J. Dines “Phonological Knowledge Guided HMM State Mapping for Cross-Lingual Speaker Adaptation”
- M. Wester and H. Liang “Cross-Lingual Speaker Discrimination Using Natural and Synthetic Speech”
- M. Shannon, H, Zen, W. Byrne “The Effect of Using Normalized Models in Statistical Speech Synthesis”

Masters theses completed within EMIME:

- Sayaka Shiota, 2009, Nagoya Institute of Technology
- Takuya Yoda, 2009, Nagoya Institute of Technology
- Kyohei Nagao, 2009, Nagoya Institute of Technology
- Zipeng Yu, 2009, Nagoya Institute of Technology
- Kaori Yutani, 2010, Nagoya Institute of Technology
- Reima Karhila, 2010, Aalto University
- Sandra Andraszewicz, 2010, University of Edinburgh
- Kyosuke Kazumi, 2011, Nagoya Institute of Technology
- Shinji Takaki, 2011, Nagoya Institute of Technology
- Toyohiro Hayashi, 2011, Nagoya Institute of Technology
- Toshinori Fukuta, 2011, Nagoya Institute of Technology
- Nagaaki Yokoyama, 2011, Nagoya Institute of Technology
- Peter Smit, expected submission 2011, Aalto University

Doctoral theses completed or largely conducted within EMIME:

- Teemu Hirsimäki, 2009, Aalto University
- Ryuta Terashima, 2010, Nagoya Institute of Technology
- Keiichiro Oura, 2010, Nagoya Institute of Technology

- Kei Hashimoto, 2011, Nagoya Institute of Technology
- Janne Pylkkönen, expected submission 2011, Aalto University
- Matt Shannon, expected submission 2011, University of Cambridge
- Sayaka Shiota, expected submission 2011, Nagoya Institute of Technology
- Hui Liang, expected submission summer 2012, Idiap / EPFL
- Lakshmi Saheer, expected submission summer 2012, Idiap / EPFL

## 4.2 – A.2 List of all dissemination activities

In addition to conference presentations, EMIME work has been included in the following presentations:

- Presentation by King on "Synthetic speech - beyond mere intelligibility" at IRCAM, Paris, France, June 2011
- Presentation by King on "Synthetic speech - beyond mere intelligibility" at the International Workshop on Voice and Speech Processing in Social Interactions, Glasgow, UK, April 2011.
- Presentation by King at University of Edinburgh, UK, March 2011
- Presentation by Kurimo at Finnish National broadcast company YLE, Finland, March 2011
- Presentation by Kurimo at L.M.Ericsson, Kirkkonummi, Finland, March 2011
- Presentation by Kurimo for students at Aalto University, Finland, March 2011
- Presentation by Kurimo for Tekniikan Maailma journal, Finland, February 2011
- Presentation by Kurimo at University of Helsinki, Finland, February 2011
- Presentation by J. Yamagishi on "New and emerging applications of 'adaptive' speech synthesis" at Cambridge statistical speech synthesis seminar series, University of Cambridge, UK, February 2011
- Presentation by Kurimo for Russian TV station Vesti, Finland, January 2011
- Presentation by M. Shannon and H. Zen on "Modelling trajectories in statistical speech synthesis" at Cambridge statistical speech synthesis seminar series, University of Cambridge, UK, January 2011.
- Invited tutorial by King on "Statistical parametric speech synthesis with some new and emerging applications" at Workshop on Image and Speech Processing International Institute of Information Technology, Hyderabad, India, December 2010
- Presentation by Tokuda, National Institute of Informatics, Japan, December 2010.
- Presentation by Kurimo at Aalto University, Finland, December 2010
- Presentation by Kurimo at press conference in Aalto University, Finland, December 2010
- Presentation by Kurimo for visitors at Aalto University, Finland, November 2010
- Invited tutorial by Yamagishi and King on "New and emerging applications of speech synthesis" at the International Symposium on Chinese Spoken Language Processing, Tainan, Taiwan, November 2010
- Presentation by Tokuda on "Speech synthesis as a machine learning problem" at Oriental COCOSDA 2010, Kathmandu, Nepal, November 2010.
- Presentation by Lee, Nara Institute of Science and Technology (NAIST), Japan, November 2010.
- Presentation by Lee, Toyohashi University of Technology, Japan, November 2010.
- Presentation by King at University of Edinburgh, UK, October 2010
- Presentation by Nankaku on "Bayesian speech synthesis framework integrating training and synthesis processes," The Chinese Academy of Science (CAS), Beijing, China, October 2010.

- Presentation by Nankaku on "Bayesian speech synthesis framework integrating training and synthesis," Microsoft Research Asia (MSRA), Beijing, China, October 2010.
- Presentation by J. Yamagishi on "New and emerging applications of speech synthesis", Nokia Research Center, Beijing, China, September 2010.
- Presentations by Kurimo at Nokia Research Center, Beijing, September 2010
- Presentation by J. Yamagishi on "1000s voices and attractive voices for speech synthesis," Signal processing laboratory (Aholab), University of the Basque country, Bilbao, Spain, September 2010.
- Presentation by J. Yamagishi on "Open-source/creative commons speech databases for speech synthesis", Special session "Open source initiative for speech synthesis", 7th ISCA Speech Synthesis Workshop, Kyoto Japan, September 2010.
- Presentations by Kurimo and Renals at NTT Labs, Kyoto, September 2010
- Presentation by Kurimo at Nagoya Insitute of Technology, September 2010
- Presentation by Kurimo at Morpho Challenge Workshop, Espoo, Finland, September 2010
- Presentation by Lee on "Speech algorithm for LVCSR," Kyoto University, Japan, August 2010.
- Presentation by Kurimo for Finnish Swedish National radio station FSR, Finland, July 2010
- Presentation by Kurimo at Sigmorphon Workshop at ACL, Uppsala, July 2010
- Presentation by Kurimo for Swedish National television station SVT1, July 2010 Master's Thesis presentation by Karhila at Aalto University, Finland, April 2010
- Presentation by Kurimo for sign language interpreter students, Helsinki, Finland, March 2010
- Presentation by Kurimo for students at Aalto University, Finland, March 2010
- Presentation by Kurimo on "Speech recognition and adaptation for multilingual and multimodal interaction" at Aalto University, Finland, February 2010.
- Presentation by Kurimo on "Decomposition of words for speech recognition, retrieval and translation" at "XXVI Fonetiikan päivät", Ilomantsi, Finland, February 2010.
- Presentation by Tokuda, Nara Institute of Science and Technology (NAIST), Japan, January 2010.
- Presentation by Kurimo on "Adaptive algorithms and Speech recognition" at University of Helsinki, Finland, January 2010.
- Presentation by Nankaku, NTT Communication Science Laboratories, Japan, January 2010.
- Presentation by Kurimo on "Speech-to-text" at "Otaniemi Demo House 2010", at Aalto University, Finland, January 2010.
- Presentation by Kurimo at Helsinki University of Technology, Finland, December 2009.
- Presentation by Kurimo on "Decomposition of words for speech recognition, retrieval and translation" at Bogazici University, Istanbul, Turkey, December 2009.
- Presentation by Tokuda, Naogya Institute of Technology (NIT), Japan, November 2009.
- Presentation by Lee on "Recent progress of speech recognition engine and decoding algorithm," Kyoto University, Japan, November 2009.
- Public lecture by King on "A survey of speech technology," Universiti Teknologi Malaysia, November 2009
- Two week course by King on speech recognition and speech synthesis at Universiti Teknologi Malaysia, November 2009
- Presentation by Tokuda on "Tutorial: Fundamentals and recent advances in HMM-based speech synthesis" at Interspeech 2009, Brighton, UK, September 2009.

- Presentation by Oura at the first HTS meeting, Brighton, UK, September 2009.
- Presentation by Kurimo on "Overview of Morpho Challenge task at CLEF 2009" at "CLEF 2009 Workshop", Corfu, Greece, September 2009.
- Presentation by Lee on "Speech algorithm for LVCSR," Kyoto University, Japan, August 2009.
- Presentation by King on speech synthesis at the MATCH summer school, Edinburgh, May 2009
- Presentation by Tokuda, Advanced Industrial Science and Technology (AIST), Japan, May 2009.
- Presentation by King on "Unsupervised adaptation for HMM-based speech synthesis," University of Science & Technology China, April 2009
- Presentation by Yamagishi on "Recent developments and applications of speaker-adaptive HMM-based speech synthesis," University of Science & Technology China, April 2009
- Presentation by King on "Hidden Markov Model-Based Speech Synthesis" at "Unified models for speech recognition and synthesis," Birmingham, UK, March 2009
- Presentation by Yamagishi on "Hundreds of Voices for HMM-based Speech Synthesis Building TTS Systems on ASR Corpora" at "Unified models for speech recognition and synthesis," Birmingham, UK, March 2009
- Presentation by Dines on "Measuring the Gap between HMM-based speech synthesis and speech recognition" at "Unified models for speech recognition and synthesis," Birmingham, UK, March 2009
- Presentations about HMM-based speech synthesis and EMIME by Yamagishi & King at iFLYTEK and the University of Science & Technology of China, Hefei, China, April 2009
- Presentation by Tokuda. KDDI lab., Japan, March 16, 2009
- Presentation by Tokuda, Nagoya Institute of Technology, Japan, February 6, 2009.
- Presentation by Tokuda, Meijo University, Japan, November 29, 2008.
- Presentation by Tokuda on "Fundamentals and recent advances in HMM-based speech synthesis," at the University of Science and Technology of China, October 30, 2008.
- Presentation by King on "Unsupervised adaptation for HMM-based speech synthesis," at Universidad de Chile, October 2008
- Presentation by Keiichi Tokuda at Nagoya Institute of Technology in Japan, September 30, 2008.
- Presentation by Tokuda on "Japan-Finland International Bilateral Seminar: F0 Modeling in HMM-Based Speech Synthesis,", Helsinki, Finland, August 6, 2008.
- Presentation by Nankaku on "A Bayesian approach to speech synthesis and a short introduction of recent work," at the Chinese Academy of Sciences, October 27, 2008.
- Presentation by Nankaku on "A Bayesian approach to speech synthesis and a short introduction of recent work," at Microsoft Research Asia, China, October 30, 2008.

Dissemination of EMIME via the EMIME web site, demonstrations and a workshop:

- EMIME publications, databases and foreground software etc are all publicly available via: `www.emime.org`
- `www.emime.org` also contains a Google Maps-based demonstration of speech synthesis including speaker adaptation.
- A voice cloning demo was developed which has been demonstrated at Interspeech, ACL, SSW7, UEDIN Schools outreach

- EMIME Mini Workshop 'Speech Synthesis Get-Together' held at Clare College, Cambridge, on 19 May 2010. Attendees were from Toshiba Research Europe Ltd, Phonetic Arts, and the Cambridge University Engineering Department.

EMIME has been covered in a variety of media:

- May, 2, 2008, Newspaper article (The Chunichi Shimbun): http://www.chunichi.co.jp/ The mobile speech-to-speech translation system which outputs speech with user's voice characteristic.
- May, 5, 2008,Radio interview (ZIP-FM): https://zip-fm.co.jp/index1.asp The mobile speech-to-speech translation system which outputs speech with user's voice characteristic.
- 28 September 2008. An article on EMIME and other projects appeared in Scotland on Sunday: http://scotlandonsunday.scotsman.com/scotland/See-the-future-from-tower.4535692.jp
- October, 8, 2008. EMIME in Scottish student newspaper http://www.studentnewspaper.org/index.php?option=com_content&view=article&id=133:new-translation-device-ensures-no-need-for-french-class&catid=34:news&Itemid=54
- April, 15, 2010, TV clip (Tokai Cable Channel): http://www.tokai-cable.jp/ HMM-based speech synthesis.
- 19 July 2010. Swedish radio interview with M. Kurimo http://sverigesradio.se/sida/artikel.aspx?programid=406&artikel=3859170
- FSR Finland July 2010 interview with M. Kurimo
- October, 1, 2010. Press release (FueTrek Co., Ltd.): http://www.fuetrek.co.jp/newsapp/NewsApp/uptmp/201001001.pdf Development of Japanese speech synthesis system using HMM-based speech synthesis.
- October, 5, 2010. Flyer: NIT handed out 500 flyers about EMIME in CEATEC JAPAN 2010.
- November, 8, 2010. Press release (FueTrek Co., Ltd.): http://www.fuetrek.co.jp/newsapp/NewsApp/uptmp/201011003.pdf Mobile phones using the spoken dialog system were released.
- November, 9, 2010. Press release (Nagoya Institute of Technology): http://www.sp.nitech.ac.jp/index.php?plugin=attach&refer=%A5%DB%A1%BC%A5%E0%2F%CA%F3%C6%BB&openfile=20101109_nit.pdf Mobile phones using the speech synthesis engine "hts_engine API" were released.
- November, 10, 2010. Newspaper article (The Asahi Shimbun): http://www.asahi.com/english/ Mobile phones using the speech synthesis engine "hts_engine API" were released.
- November, 10, 2010. Newspaper advertisement (The Asahi Shimbun): http://www.asahi.com/english/ Mobile phones which are controlled by a spoken dialog system were released.
- Press release Aalto University, December, 2010
- Nelonen Finland December 2010,interview with M. Kurimo
- Vesti Russia February 2011, interview with M. Kurimo http://www.vesti.ru/videos?vid=316647
- March, 28 2011. RTL Germany, interview with M. Kurimo http://research.ics.tkk.fi/mi/RTLnachtjournal-20110328.avi

## 4.2 – B Exploitable foreground

### Plan for the use and dissemination of foreground

(This section is taken from Annex I, and we have added commentary in *italics* describing how the plan was implemented in the project.)

... we propose a separate budget to cover evaluation and dissemination activities. This budget will be managed by UEDIN, with the help of the project board, and will support dissemination activities such as workshops and the organization of evaluation activities.

*This plan was executed successfully. The budget was used to assist in the organisation of the Blizzard Challenge, which was an excellent way to raise the visibility of EMIME, as well as providing us with useful external evaluations. We also supported a meeting with two key players in speech synthesis – Toshiba and Phonetic Arts – in which we presented our results and gained useful feedback.*

Dissemination to a technical audience will exploit all traditional means including:

- Participation in, and organization of, relevant technology evaluations, such as Blizzard. *This plan was executed successfully.*
- Workshop organization, possibly in conjunction with related projects, or as a satellite to well- established conferences. *This plan was executed in part, with one small workshop, as well as supporting the Blizzard Workshops*.
- Production of a final public report (deliverable D5.6), including state-of-the-art reviews, distributed to a substantial, but targeted mailing list comprising scientists, industry thought leaders, policy makers. *This report has been prepared (it is the current document!) and will be distributed according to plan.*
- Publication of scientific papers in international conferences and scientific journals, where the FP7 funding will always be duly acknowledged. Whenever possible, we will also publish joint papers presenting EMIME as a whole, in addition to papers presenting progress in specific areas. *This plan was executed successfully.*
- Publication of articles in technical journals and institutional magazines. *This plan was executed successfully.*
- The project website will include a browsable and searchable database of all public outputs and papers resulting from the project. *This plan was executed successfully.*
- Open source distribution of software, such as Festival and HTS. *This plan was executed successfully.*
- Participation in, and leveraging on, other (national) research programmes. Although all the EMIME partners are very active at the national level, EMIME will further reinforce the opportunities for leveraging on these projects, while advertising our work at the national level and attracting more interest. *This plan was executed successfully, with EMIME finding useful links with a number of other projects.*

Regarding external information aimed at the *non-technical audience*, all the partners are aware of the necessity to communicate as with the general public and of the need for public understanding of science. Personalised speech-to-speech translation, particularly as embodied in the real-time demonstrator that we will build, is a topic that is ideal for attracting public interest. The partners will work with their respective public relations and media offices to capitalise on this. There will be a specific section of the project website aimed at a general audience. All EMIME partners are active in presenting science and technology to a young audience (e.g., via science festivals or multimodal art exhibits) and the results of EMIME will further enrich our activities in these areas.

*The EMIME voice cloning demo has already been used in schools outreach at UEDIN. We hope to provide the realtime demo for the same purpose in the future.*

## 4.2 – B.1 List of applications for patents, trademarks, registered designs, etc

None. Foreground has been released as open source under the least restrictive terms possible, as described in the next section.

## 4.2 – B.2 List of exploitable foreground

### 4.2 – B.2.1 Introduction

This section describes all foreground created in the project that is being publicly released under an open source license.

### 4.2 – B.2.2 The situation regarding the STRAIGHT vocoder

STRAIGHT is proprietary software of Advanced Telecommunications Research Institute International, Japan. The conditions of the research license of STRAIGHT are

1. For models built using STRAIGHT with research only licenses, the models should have a research only caveat too.
2. We must not re-distribute STRAIGHT code or runtime binaries.

Therefore TTS models built using STRAIGHT cannot be released under an open source license such as FreeBSD. They muse be released under a specific research license. Feature extraction modules using STRAIGHT cannot be released under any circumstances.

### 4.2 – B.2.3 License of Google's translation APIs

The licenses of Google translation APIs are given at the following URL `http://research.google.com/university/translate/terms.html`, in which the clause 3.1 prohibit the use for any commercial purpose.

> 3.1 You may use the API and may use and reproduce the Results for academic research purposes only and in accordance with the Terms and any applicable guidelines provided to you by Google. You may not use the API or the Results for, as the basis of, or in connection with any commercial use, application, or development. You may not use, distribute or otherwise make available your Application for any commercial purpose.

Therefore, the EMIME Google translation module must be released under the same research license.

### 4.2 – B.2.4 UCAM bilingual dataset

**URL:** `http://www.emime.org/participate/ucam-bilingual-database`

This dataset contains the speech of four male non-native speakers of English. In the case of the European languages considered (French, Italian and Dutch) the speech corresponds to utterances selected from the Europarl corpus of parallel text of European parliament proceedings. In the case of the Mandarin speaker, the speech corresponds to a subset of the NIST 2008 Chinese-English MT evaluation parallel texts. Each speaker provided speech in his native language as well as the parallel translated speech in English. The speech was recorded using a Sennheiser close-talking microphone in a quiet office at a sampling rate of 44.1 kHz.

Table 4.2 – B.2.4 displays the contents of the dataset. The first column records the speaker native language, while the second column displays the number of parallel utterances (i.e. in both his native language and English) recorded. Each speaker delivers an additional set of 40 English-only utterances which were used as evaluation data in cross-lingual adaptation experiments.

| Native language | # Parallel utterances | # English only utterances |
|---|---|---|
| French | 130 | 40 |
| Italian | 130 | 40 |
| Dutch | 130 | 40 |
| Mandarin | 89 | 40 |

Table 4.2 – B.2a: Contents of the UCAM bilingual dataset.

### 4.2 – B.2.5 UEDIN bilingual dataset

The EMIME Bilingual Finnish/English German/English Database
**URL:** http://www.emime.org/participate/emime-bilingual-database
This dataset contains the speech of 28 speakers. Seven male and seven female speakers of Finnish and seven male and seven female speakers of German were recorded. Each speaker read sentences in English and in their native language. The English, Finnish and German prompt sets each contain 25 Europarl sentences, 100 news sentences and 20 semantically unpredictable sentences (SUS). The 25 Europarl sentences were selected from the ACL WMT 2008 test set of the Europarl (proceedings of the European Parliament) parallel corpus. The news sentences for English were taken from the Wall Street Journal 1 corpus, comprising 40 enrolment sentences and 60 test set sentences. The Finnish sentences were selected from the Speecon corpus. German news sentences were selected from the test set portion of German Globalphone. The database includes a more detailed README.

The EMIME Bilingual Mandarin/English Database
**URL:** http://www.emime.org/participate/emime-bilingual-database
This dataset contains the speech of 14 speakers. Seven male and seven female speakers were recorded reading Mandarin and English sentences. The English and Mandarin prompts sets each contain 25 Europarl sentences (for Mandarin these sentences were translated from English), 100 English news sentences(same as above) and 124 Mandarin news sentences (selected from the Speecon corpus), respectively, and 20 semantically unpredictable sentences (SUS). The database includes a more detailed README.

### 4.2 – B.2.6 Nokia Bilingual Database for Realtime-Demo

**URL:** http://www.emime.org/participate/emime-speakeradaptationdata-8khz
Speaker adaptation data for EMIME Real-time demo These data are recorded by Nokia mobile phone N97 mini. The sample rate of the speech is 8KHz and the speech are transmitted from mobile phones to Linux server with AMR codec. This dataset contains the speech of 6 native Mandarin speakers and 1 native US-English speaker. 4 male and 2 female Mandarin speakers are recorded. Each speaker read sentences in Mandarin and English. 100 Chn/Eng parallel utterances from Nokia Phrasebook database are selected as part I. 100 Chn/Eng non-parallel utterances from SMS database as part II. For the native US-English speaker, only 100 English sentences from Phrasebook database are recorded.

### 4.2 – B.2.7 Machine translation lattices

**URL:** http://data.cstr.ed.ac.uk/emime/ch-en.tar.gz
**URL:** http://data.cstr.ed.ac.uk/emime/fi-en.tar.gz
**URL:** http://data.cstr.ed.ac.uk/emime/mt-lattices.tar.gz
Lattices corresponding to Finnish-English translations (Europarl dataset, last quarter of the year 2000) and Mandarin-English translations (NIST MT08 evaluation dataset). The system which generated the Finnish-English

lattices is described in 'A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. In Procedings of NAACL-HLT, 2009.'

### 4.2 – B.2.8 Tools

HTS
**URL:** `http://hts.sp.nitech.ac.jp/archives/2.2beta/HTS-2.2beta_for_HTK-3.4.1.tar.bz2`
Since December 2002, an open-source software toolkit named "HMM-based Speech Synthesis System (HTS)" is publicly released to provide a research and development platform for statistical parametric speech synthesis by NIT. Various organizations include EMIME project currently use it to conduct their own research projects. Latest version (ver. 2.2) of HTS support cross-lingual speaker adaptation based on a state-level mapping learned using minimum KLD which is one of core technology of EMIME deliverable. Furthermore, next version of HTS will support VTLN technique proposed by Idiap.

Festival
**URL:** `http://www.cstr.ed.ac.uk/projects/festival/`

Autoregressive HMM for HTS
**URL:** `http://mi.eng.cam.ac.uk/research/emime/ar-for-hts`
An implementation of the autoregressive HMM built on top of HTS. Provides the ability to do embedded re-estimation and decision tree clustering for linear-Gaussian autoregressive output distributions. Public release as a patch file for HTS 2.1 under a BSD-style license.

armspeech
**URL:** `https://github.com/MattShannon/armspeech`
**URL:** `http://mi.eng.cam.ac.uk/research/emime/armspeech`
*armspeech* is a set of tools for autoregressive acoustic modelling. It provides a framework and example experiments for investigation of non-linear autoregressive acoustic models, including Gaussian process and mixture-of-experts acoustic models. It is written in python. Public release under a BSD-style license.

Two-pass decision tree construction
**URL:** `http://data.cstr.ed.ac.uk/emime/decision-tree-software.tar.gz`
An extension of the HTS HHEd tool which provides incremental decision tree growing functionality developed at UCAM.

EM VTLN adaptation
**URL:** `http://www.emime.org/participate/tools/em-vtln-adaptation`
This allows VTLN adaptation within the context of HTS. This code will include the possibility to perform VTLN adaptation as a global warping of the spectrum using base classes and also as multiple warping parameters for different phoneme classes using regression trees (similar to CMLLR adaptation). The code will be incorporated into the mainline HTS distribution (HTS-2.2, alluded to above) from NITech.

VTLN as CSMAPLR prior
**URL:** `TBC`
This code will allow VTLN to be used in combination with CMLLR adaptation by using a VTLN transformation matrix as a prior to the CSMAPLR transformation. This technique can be a global VTLN prior matrix in the structural MAP adaptation or multiple VTLN matrices as priors to the different nodes in the regression tree during CMLLR adaptation. The first option has been developed and showed good performance improvements. As of April 2011 the second option is currently still being developed. However, when complete it will be released in the

same manner as the EM VTLN implementation above.

Java HMM editor
**URL:** http://www.emime.org/participate/tools/java-hmm-editor
This is a program written in java to perform decision tree marginalisation operations on HTK HMM files. It will be released as a package with the EMIME software.

Cross language state mapping
**URL:** http://www.emime.org/participate/tools/cross-language-state-mapping
A set of programs written mainly in perl to do cross language state mapping and adaptation. It will be released as a package with the EMIME software.

WSJ training scripts
**URL:** http://www.emime.org/participate/tools/wsj-training-scripts
A set of scripts to train ASR models. They are based on HTS scripts from Junichi Yamagishi, but modified to build ASR models. They were used for the experiments in the "Measuring the gap" paper.

### 4.2 – B.2.9 Framework of Research-Demo

**URL:** http://www.emime.org/participate/tools/the-framework-of-emime-research-demo
The research demonstrator enables both individual component and integrated system evaluation and supports technical exploration of the research questions.

In general, the research framework has file-based structure and the modules interact using files, through which the modules share data.

the structure of the scripts in research demonstrator can be described as comprising three levels: root level scripts control and start up the whole system, module level scripts control and configure the module specifics, and instance level scripts run specific processing in modules with configuration information transferred from the root and modified at the module level. The scripts are written in Perl or tcsh.

Please refer D4.5 for details of the Research-Demo framework.

### 4.2 – B.2.10 Framework and Software of Realtime-Demo

Framework and Server end application
**URL:** http://www.emime.org/participate/tools/framework-of-emime-realtime-demo
The realtime demo framework include the kernel of research demo and some supporting scripts and application to connect mobile phone client and linux server.

"EMIMEServer" is an ACE based server end application for Realtime demo. It processes the interaction between clients and the personalized speech to speech translation core engine in linux server.

The supporting scripts are called by "EMIMEServer" and initialize the scripts in research demo kernel.

Client software
**URL:** http://www.emime.org/participate/tools/client-software-of-emime-realtime-demo
"EMIME" mobile application is implemented in Symbian S60 5th mobile environment. It is the client end software of EMIME Realtime-Demo. The "EMIME" records speech, shows recognized/translated text and plays output speeches of EMIME Realtime-Demo. Also, from this client software, users can edit the recognized text to rectify recognition errors and select whether to do adaptation or not in server.

### 4.2 – B.2.11 TTS average voice models

Chinese
**URL:** http://www.emime.org/participate/tts-average-voice-models/chinese
Chinese Speecon Average Voice Model

English
**URL:** http://www.emime.org/participate/tts-average-voice-models/english_si284
WSJ0 + WSJ1 English Average Voice Model
**URL:** http://www.emime.org/participate/tts-average-voice-models/english_si84
WSJ0 English Average Voice Model
**URL:** http://www.emime.org/participate/tts-average-voice-models/english_wsjcam
WSJCAM0 Average Voice Model

Finnish
**URL:** http://www.emime.org/participate/tts-average-voice-models/finnish
Finnish Speecon Average Voice Model

Japanese
**URL:** http://www.emime.org/participate/tts-average-voice-models/japanese
Japanese JNAS Average Voice Model

### 4.2 – B.2.12 ASR models

Japanese
**URL:** http://data.cstr.ed.ac.uk/emime/Japanese_ASR_models.tar.gz

Japanese News paper Article Sentences (JNAS) database were used for training HMMs. To train an acoustic model for speech recognition was trained from 37k sentences uttered by 122 male and 122 female speakers. Speech signals were sampled at a rate of 16 kHz and windowed by a 25ms Hamming window with a 10ms shift. The feature vectors consisted of 13-dimension mel-cepstral coefficients plus the zero-th coefficient, their dynamic, and acceleration coefficients. The topology is three state left-to-right HMMs. Additionally, the language models used for speech recognition were based on Mainichi news paper corpus.

English
**URL:** http://www.emime.org/participate/asr-models/english-asr-models

Models are trained on the si-84 subset of the LDC WSJ0 corpus. Both MFCC and PLP models are provided, with two phonetic representations (gam and arpabet) based on the Unilex dictionary. In addition, the MFCC arpabet models are provided with the AMR codec.

The release is subject to the LDC conditions; that is, the acoustic models are released as a derived work, but for research only. The language model is a part of the WSJ0 corpus and can only be released to parties who can show they have permission to use the WSJ0 corpus.

### 4.2 – B.2.13 ASR models for the Realtime Demo

**URL:** TBC

The ASR models for the Realtime demo are all trained with speech data that was encoded and decoded with the AMR codec to simulate the conditions of the demo setup. Also, care was taken that the models were of such size that the performance would be reasonable, up to Realtime performance.

The main ASR engine used for the Realtime demo was Juicer, which uses Finite State Models in recognition. The accoustic models are trained with the HTK toolkit.

Models for the Realtime demo are the AMR versions of the ones stored in the Research demo.

Finnish - Phrasebook - phrasebook_speecon_amr.tar.gz
**URL:** http://data.cstr.ed.ac.uk/emime/phrasebook_speecon_amr.tar.gz
This model was created by Aalto. It is build according to the description in Chapter 3.4.1 of EMIME Deliverable 4.6. This model is distributed under a FreeBSD license. (Note: this item is in the process of being uploaded to the above URL.)

Finnish - Ngram - vari_10g_speecon_amr.tar.gz
**URL:** http://data.cstr.ed.ac.uk/emime/vari_10g_speecon_amr.tar.gz
This model was created by Aalto. The accoustic part is made according to the description in Chapter 3.4.1 of EMIME Deliverable 4.6. The language model is a 15k morph-based ngram model trained on the Kielipankki corpus. This model is distributed under a FreeBSD license. (Note: this item is in the process of being uploaded to the above URL.)

English - Phrasebook - si284.cmu.AMR.phrasebook.tar.gz
**URL:** http://data.cstr.ed.ac.uk/emime/si284.cmu.AMR.phrasebook.tar.gz
This model was created by Aalto. It is trained with the Wall Street Journal SI-284 corpus and the language model is a direct translation of the Finnish Phrasebook language model. The WSJ corpus is licensed by LDC, therefore the distribution restrictions mentioned in the section "LDC data" apply.

Mandarin - Phrasebook - Mandarin_PCOM_phrasebook.tar.gz
**URL:** http://www.emime.org/participate/asr-models/mandarin-fsts-based-on-phrasebook-for-juicer-engine
This model was created by Nokia. It is trained with PCOM corpus and the language model is a direct translation of the Finnish Phrasebook language model. It can be used for research only.

Additional comments on how foreground my be exploited, and by whom, can be found in Section 4.1.4.

# 4.3

# Report on societal implications

## A General Information

| |
|---|
| Title of Project: Effective Multilingual Interaction in Mobile Environments |
| Name and Title of Coordinator: Prof. Simon King, University of Edinburgh |

## B Ethics

| | |
|---|---|
| **1. Did your project undergo an Ethics Review (and/or Screening)?** | No |
| **2. Please indicate whether your project involved any of the following issues (tick box) :** | |
| **Research on humans** | |
| Did the project involve children? | |
| Did the project involve patients? | |
| Did the project involve persons not able to give consent? | |
| Did the project involve adult healthy volunteers? | YES |
| Did the project involve Human Genetic Material? | |
| Did the project involve Human biological samples? | |
| Did the project involve Human data collection? | |
| **Research on Human embryo/foetus** | |
| Did the project involve Human Embryos? | |
| Did the project involve Human Foetal Tissue / Cells? | |
| Did the project involve Human Embryonic Stem Cells (hESCs)? | |
| Did the project on human Embryonic Stem Cells involve cells in culture? | |
| Did the project on human Embryonic Stem Cells involve derivation of cells from Embryos? | |
| **Privacy** | |
| Did the project involve processing of genetic information or personal data (e.g.,. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction) | |
| Did the project involve tracking the location or observation of people? | |
| **Research on Animals** | |
| Did the project involve research on animals? | |
| Were those animals transgenic small laboratory animals? | |
| Were those animals transgenic farm animals? | |
| Were those animals cloned farm animals? | |
| Were those animals non-human primates? | |
| **Research Involving Developing Countries** | |
| Did the project involve the use of local resources (genetic, animal, plant etc)? | |
| Was the project of benefit to local community (capacity building, access to healthcare, education etc)? | |
| **Dual Use** | |
| Research having direct military application | |
| Research having the potential for terrorist abuse | |

# C Workforce Statistics

**3. Workforce statistics for the project: Please indicate in the table below the number of people who worked on the project (on a headcount basis).**

| Type of Position | Number of Women | Number of Men |
|---|---|---|
| Scientific Coordinator | 1 | 5 |
| Work package leaders | 0 | 5 |
| Experienced researchers (i.e. PhD holders) | 2 | 17 |
| PhD Students | 3 | 9 |
| Other | 3 | 16 |
| **4. How many additional researchers (in companies and universities) were recruited specifically for this project?** | | 5 |
| Of which, indicate the number of men: | | 5 |

# D Gender Aspects

| | | | | | |
|---|---|---|---|---|---|
| **5. Did you carry out specific Gender Equality Actions under the project?** | Yes (Nokia only) | | | | |
| **6. Which of the following actions did you carry out and how effective were they?** | | | | | |
| | Not at all effective —- Very effective | | | | |
| Design and implement an equal opportunity policy | o | o | o | o | x |
| Set targets to achieve a gender balance in the workforce | o | o | o | o | x |
| Organise conferences and workshops on gender | o | o | o | x | o |
| Actions to improve work-life balance | o | o | o | o | x |
| Other: | none | | | | |
| **7. Was there a gender dimension associated with the research content i.e. wherever people were the focus of the research as, for example, consumers, users, patients or in trials, was the issue of gender considered and addressed?** | No | | | | |

# E Synergies with Science Education

**8. Did your project involve working with students and/or school pupils (e.g. open days, participation in science festivals and events, prizes/competitions or joint projects)?**

Yes. The voice cloning demo is being used for school outreach and science eduction at UEDIN

| | |
|---|---|
| **9. Did the project generate any science education material (e.g. kits, websites, explanatory booklets, DVDs)?** | No |

# F Interdisciplinarity

**10. Which disciplines (see list below) are involved in your project?**

| | |
|---|---|
| Main discipline : | 1.1 Mathematics and computer sciences |
| Associated discipline : | 5.4 Other social sciences – Linguistics |

# G Engaging with Civil society and policy makers

| | |
|---|---|
| Did your project engage with societal actors beyond the research community? | No |

# H Use and dissemination

| | |
|---|---|
| 14. How many Articles were published/accepted for publication in peer-reviewed journals? (conference papers are not included, of which there are around 65) | 12 |
| To how many of these is open access provided? | 6 |
| How many of these are published in open access journals? | 0 |
| How many of these are published in open repositories? | 0 |

| | |
|---|---|
| To how many of these is open access not provided? | 6 |
| Please check all applicable reasons for not providing open access: | |
| publisher's licensing agreement would not permit publishing in a repository | |
| no suitable repository available | |
| no suitable open access journal available | |
| no funds available to publish in an open access journal | |
| lack of time and resources | Yes |
| lack of information on open access | |
| other | |
| 15. How many new patent applications (priority filings) have been made? ("Technologically unique": multiple applications for the same invention in different jurisdictions should be counted as just one application of grant). | 0 |
| 16.Indicate how many of the following Intellectual Property Rights were applied for (give number in each box). | |
| Trademark | 0 |
| Registered design | 0 |
| Other | 0 |
| 17. How many spin-off companies were created / are planned as a direct result of the project? | 0 |
| 18. Please indicate whether your project has a potential impact on employment, in comparison with the situation before your project: | Difficult to estimate / not possible to quantify |
| 19. For your project partnership please estimate the employment effect resulting directly from your participation in Full Time Equivalent (FTE = one person working fulltime for a year) jobs: | 45 (total project effort of 541 person months) |

# I Media and Communication to the general public

| | |
|---|---|
| 20. As part of the project, were any of the beneficiaries professionals in communication or media relations? | No |
| 21. As part of the project, have any beneficiaries received professional media / communication training / advice to improve communication with the general public? | Yes |
| 22 Which of the following have been used to communicate information about your project to the general public, or have resulted from your project? | |
| Press Release | Yes |
| Coverage in specialist press | Yes |
| Media briefing | |
| Coverage in general (non-specialist) press | Yes |
| TV coverage / report | Yes |
| Coverage in national press | Yes |
| Radio coverage / report | Yes |
| Coverage in international press | |
| Brochures /posters / flyers | Yes |
| Website for the general public / internet | |
| DVD /Film /Multimedia | |
| Event targeting general public (festival, conference, exhibition, science cafe) | Yes |
| 23 In which languages are the information products for the general public produced? – we have included general (non-specialist) media coverage under this heading | English, Japanese, Swedish, Finnish |