An open source speech synthesis module for a visual-speech recognition system

Sotiris Manitsaris, Bruce Denby, Florent Xavier, Jun Cai, Maureen Stone, Pierre Roussel, Gérard Dreyfus

ACOUSTICS 2012

An open source speech synthesis module for a visual-speech recognition system

S. Manitsaris[a], B. Denby[a], F. Xavier[b], J. Cai[a], M. Stone[c], P. Roussel[a] and G. Dreyfus[a]

[a]Laboratoire Signaux, Modèles et Apprentissage Statistique, 10 rue Vauquelin, 75231 Paris cedex 05
[b]Laboratoire traitement et communication de l'information, 46 Rue Barrault 75013 Paris
[c]University of Maryland, University of Maryland, College Park, MD 20742, USA
sotiris.manitsaris@espci.fr

A Silent Speech Interface (SSI) is a voice replacement technology that permits speech communication without vocalization. The visual-speech recognition engine of the proposed SSI is based on vocal tract imaging. The system aims to give the laryngectomised speaker the opportunity to speak with his/her original voice. The visual-speech recognition engine of the SSI outputs a text sentence, which is imported to the speech synthesis module in order to synthesize speech in French or English. This paper presents the speech synthesis module of a SSI that uses the open-source MaryTTS (Text-To-Speech). A new module of phonetic transcription has been developed and integrated into MaryTTS. In addition, English and French semi-HMM (Hidden Markov Models) model voices have been built. The SSI can be remotely controlled using a mobile device and the new voices are installed in a Web Server.

1 Introduction

This paper presents the speech synthesis module of the portable and stand alone Silent Speech Interface (SSI) that has been developed in the SIGnal processing MAchine learning (SIGMA) laboratory at the Ecole Supérieure de Physique et de Chimie Industrielle de la Ville de Paris (ESPCI ParisTech). The system performs visual-speech recognition based on the vocal tract imaging of the speaker. The recognition module exports text, which is synthesised using a voice trained on the speaker's vocal tract. The speech synthesis module is based on a Text-To-Speech system (TTS) and it is available for French and English.

SSIs are systems intended to recognize and/or synthesize speech based on sensor data collected from the articulators, in particular when glottal activity is absent [1]. Development of SSIs, using a wide variety of sensor types, remains an active area of research. Real time ultrasound (US) and video imaging of the vocal tract was shown to be effective for offline, visio-phonetic continuous speech recognition, in a fixed, "benchtop" SSI [2, 3]. Though interesting as a proof of principle, such a system is impractical for everyday applications since both the SSI and the user are required to remain immobile. A system employing a professional ultrasound acquisition helmet, with an added camera, was described in [4]. Although a first step towards a portable US SSI, the instrumentation used in that test proved too cumbersome for prolonged use, required a controlled acquisition environment (*lighting, etc.*), and ultimately remained an offline tool only.

TTS converts a given text into an audio speech signal. Thus, many different useful applications of this technology can be considered. Lets consider for example a phone book that contains thousands - even millions - of entries. It is not possible for a speaker to record all these phone numbers but it is possible for a TTS to synthetise a voice reading them. The project GRETA [5] aims to create conversational agents with the help of the commercial system AcapelaTTS [6]. Within this context, speech synthesis can be used as a daily supplement for human-human communication aspects. The REVOIX project proposes a visual-speech recognition system based on vocal tract imaging and speech synthesis for laryngectomised speakers [7].

2 Overview of the portable SSI

The SSI of the SIGMA laboratory is based on the multimodal data acquisition of ultrasound and optical image sequences for tongue and lips respectively. This technology is effective for online, visio-phonetic continuous speech recognition. The ultrasound probe and the optical camera of the SSI have been mounted on a new practical and easily carried helmet, which can be fit into a small carrying case and it can be operated by the speaker in any environment (Figure 1). This SSI goes with the "Ultraspeech" software package [8] that offers multithread programming for the synchronous acquisition of video streams at 60 fps along with the audio. The system is completely portable since it uses a portable battery. Mobile devices, like smartphones or pads, can remotely control the SSI.

The system implements a visual-speech recognition and synthesis methodology. The first step of this methodology is the importation of both the ultrasound and optical image sequences into the recognition as well as the image pre-processing procedure that contains the resizing of the images into 64x64. The feature extraction is based on PCA

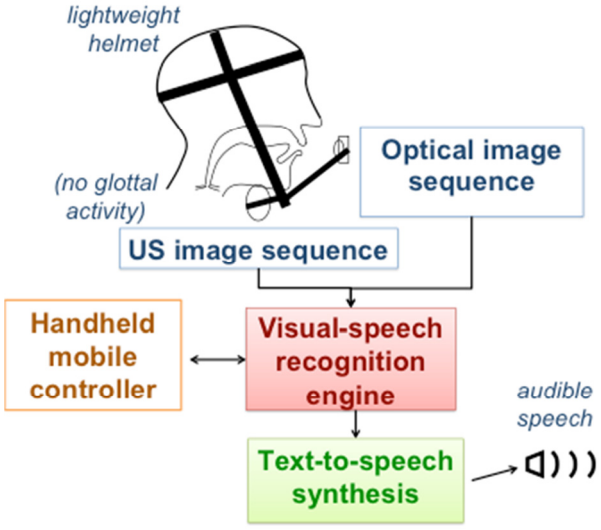

Figure 1: Architecture of the portable Silent Speech Interface.

or DCT techniques, which are applied on the images. These features are used to train several levels of Hidden Markov Models (HMM) using the HTK toolkit [9]. Language models are used (a) for English: "WSJ0", "Gigaword", "PhoneDial" and "Everyday English" and (b) for French: the Polyvar-based language model and "Everyday French". The trained models contribute to the visual-speech recognition engine with the help of Julius (*Triphones, Bigrams/Trigrams*) [10]. The recognition engine exports text sentences in French or English. This text is then synthesised using the speech synthesis module that is based on the MaryTTS system. The TTS can build semi-HMM voices in French and English and its architecture can be either standalone or client/server.

3 The speech synthesis module

The speech synthesis module of this SSI is based on the MaryTTS system (Modular Architecture for Research on speech sYnthesis) [11]. MaryTTS is an open-source and

multilingual TTS platform written in Java. It was originally developed as a collaborative project of DFKI's Language Technology lab (Deutsche Forschungszentrum für Künstliche Intelligenz) and the Institute of Phonetics at Saarland University in Germany and is now being maintained by DFKI. MaryTTS supports German, British and American English, Telugu, Turkish, and Russian. It comes with toolkits for quickly adding support for new languages and for building unit selection and HMM-based synthesis voices. This TTS is fully modular and open and it allows the simple integration of a new language, as long as a module of its phonetic transcription is created. A tool for the creation of new voices is also available. So, in order to build new French and English voices for speech synthesis, there was no need to develop a new dedicated system. For English, all the required rules for Natural Language Processing (NLP) were already available for MaryTTS and a new module for speech synthesis in French together with the NLP rules were integrated (see Figure 2).

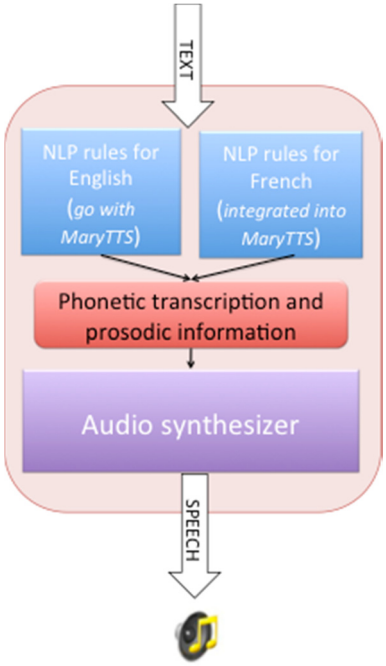

Figure 2: The speech synthesis module of the SSI.

The HMM-based speech synthesis is concatenated and it consists of two main phases: a) the training and b) the synthesis phase. The training phase allows the acoustic feature extraction (*fundamental frequency, coefficients MFCC etc.*) from a recorded text corpus. These features are modeled with the help of HMMs. The synthesis phase allows the concatenation of the best items of a phonetized text. The main advantage of such a system is the quality and the intelligibility of the speech signal, but also the similarity between the synthesized voice and the speaker's original voice. The acoustic parameters of the speaker are modeled during the training phase.

In order to integrate French into MaryTTS, a module for the phonetic transcription has been developed within the cooperation framework between SIGMA (ESPCI ParisTech) and LTCI (Télécom ParisTech) laboratories. This module is based on the grapheme-phoneme system LIA_phon, created by Frédéric Bechet [12]. The system is actually able to extract the phonetic transcription of a plain text in French (SAMPA format). The modeling of the prosody is based on a generic module, which is applied to almost all languages and is based on the ToBI standard [13].

It should also be mentioned that in order to create the voice using the built-in tool, an appropriate text corpus is prerequisite. For the needs of our research the Polyvar corpus for French and the CMU Arctic for English have been chosen. One male English and two male French voices have been built. The Polyvar corpus initially contained 3000 sentences. For the needs of the REVOIX project, two scenarios have been tested for the creation of French voices: (a) the use of the first 2000 sentences of the Poylvar corpus and (b) the use of the first 1000 sentences of a modified version of the Polyvar corpus using a greedy algorithm in order to select the most phonetically rich sentences.

4 Natural Language Processing for French

The NLP components are a prerequisite in order to create French voices for MaryTTS. In other words, a new module that takes as input RAWMARYXML and outputs PHONEMES had to be developed and integrated into MaryTTS [14, 15] (Figure 3). As a first attempt, a new complete gapheme to phoneme system together with preprocessing modules inspired from the German language have been created and tested. These preprocessing modules were used to clean the text of dates, times, phone numbers,

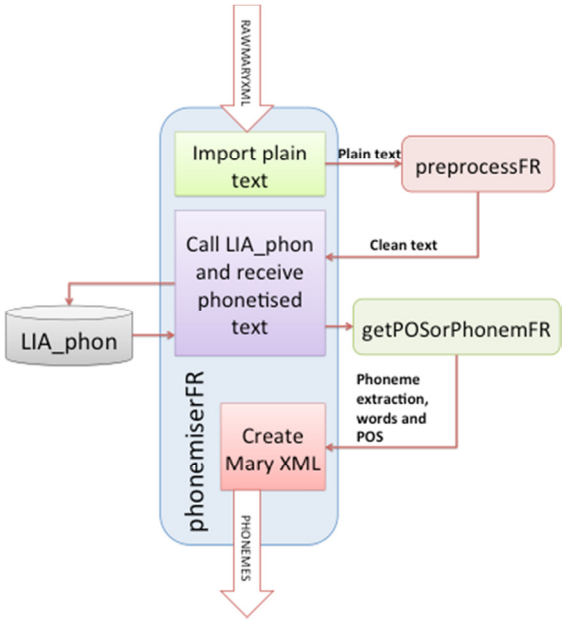

Figure 3: NLP components for French are integrated into MaryTTS.

measurement units, currencies and abbreviations. However, due to technical constraints, it was impossible to continue in this way. Therefore, LIA_phon was chosen to perform a complete phonetic transcription since it was available for a research use (*GNU license*). In order to call LIA_phon from Java software, system call techniques were used. Thus, the "phonetizer" for French consists of three classes: (a) the *PhonemiserFR*, which can be considered as *main* class. *PhonemiserFR* gets the RAWMARYXML from TextToMaryXML, calls LIA_phon and creates MaryXML

and the output PHONEMES; (b) the *getPOSorPhonemFR*, which adapts the output of LIA_phon to the MaryTTS format using the phonetical transcription in SAMPA and the morphosyntactic labels and finally (3) the *preprocessFR*, which applies a complement to LIA_phon phonetic transcription. These classes do not use very sophisticated programming. Nevertheless, they are crucial for the proper functioning of the system in general. For example, a mistake like a / _ / at the beginning of the phonetic transcription can distort the rest of the voice creation process. The creation of the NLP components for French is aligned with the requirements of the client/server architecture of MaryTTS. Each class is *thread-safe*. That means that it can be instantiated as many times as possible (*that is to say, executed by multiple processes*) without problems.

5 French and English corpora

The Polyvar corpus has been used for the creation of two male voices [16]. Initially, Polyvar consisted of 31040 French sentences. Repeated sentences have been removed and the comments have been deleted. So, 11669 short sentences in total were selected. The selection criterion is that sentence length must be between 60 and 110 characters including spaces. These 11669 short sentences contain 15119 triphones in total. By using a greedy algorithm, 3000 sentences have been selected in order to build the French visual-speech training corpus. These 3000 sentences contain 14638 triphones. In the modified version of Polyvar, the sentences are classified by number of triphones. Thus, we expect to find the most phonetically rich sentences at the beginning of the modified corpus. For the speech synthesis, one male voice has been created using the first 1000 sentences of the modified Polyvar corpus from the greedy algorithm and a second male voice using the first 2000 sentences of the initial Polyvar corpus.

The CMU Arctic corpus has been used for training MaryTTS with a new English voice [17]. CMU Arctic consists of 1132 sentences (79.6% of biphones and 13.7% of triphones). These sentences have been recorded with the speaker's voice and both text and acoustic signals have been used for the creation of the new voice with the help of MaryTTS.

6 Conclusions and future work

In this paper, an open source speech module for a Silent Speech Interface has been presented. The visual-speech recognition engine of the SSI outputs a text sentence, which is imported to the speech synthesis module in order to synthesize speech in French or English. The speech synthesis is based on the open source text-to speech system MaryTTS. All the tools to build an English voice were already available with MaryTTS. For French, the NLP rules had to be integrated into the system. The complete grapheme to phoneme and open source software LIA_phon has been used for this purpose. This speech synthesis module can either be used as a tool to create new synthesised voices based on patient's voice recordings before his operation, either as a part of the complete Silent Speech Interface (visual-speech recognition and synthesis).

Different evaluation scenarios will be studied for each of the created voices as well as new French and English voices will be built in order to propose a voice modelling protocol for patients.

Acknowledgments

We would like to thank Frédéric Bechet from the Avignon University (France) for his contribution with LIA_phon, Marc Schröder and Shatish Pammi from DFKI (Germany) for their help during the integration of the speech synthesis module into MaryTTS, Thodoris Mironidis from University of Macedonia (Greece) for his help during the installation and the configuration of the module and finally Thomas Hueber from GIPSA-Lab (France) for Ultraspeech, Catherine Pelachaud and Gérard Chollet from (Télécom ParisTech) for their expertise, advice and mentoring.

References

[1] Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M. & Brumberg, J.S. (2010), *Silent speech interfaces. Speech Communication*, v.52 n.4, p.270-287, April, 2010

[2] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (2009a). "Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasoundbased Silent Speech Interface," in *Interspeech* (Brighton, UK), pp. 640-643.

[3] Hueber T., Chollet G., Denby B., Dreyfus G. and Stone M. (2008). "Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips," in *Interspeech,* Brisbane, Australia, pp. 2028-2031.

[4] Florescu V.-M., Crevier-Buchman L., Denby B., Hueber T., Colazo-Simon A., Pillot-Loiseau C., Roussel P., Gendrot C., Quattrocchi S. (2010): "Silent vs vocalized articulation for a portable ultrasound-based silent speech interface", In *Interspeech-2010*, 450-453.

[5] Pelachaud C. "The GRETA project: Embodied Conversational Agent". Retrieved February 27, 2012 from: http://perso.telecom-paristech.fr/~pelachau/Greta/.

[6] AcapelaTTS. Retrieved February 27, 2012 from: http://www.acapela-group.com/text-to-speech-interactive-demo.html

[7] Denby B. "The REVOIX project SIGMA-ESPCI ParisTech". Retrieved February 27, 2012 from: http://www.neurones.espci.fr/Revoix/.

[8] Hueber, T., Chollet, G., Denby, B., and Stone, M. "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," *Proceedings of International Seminar on Speech Production* (Strasbourg, France), pp. 365-369. 2008.

[9] Young, S., Evermann, G., Gales, M., et al., "*The HTK Book*", Retrieved on 9 March 2012 from:

http://htk.eng.cam.ac.uk/docs/docs.shtml, accessed on 06 Mar. 2012.

[10] Lee, A., Kawahara, T. and Shikano, K., "Julius – An Open Source Real-time Large Vocabulary Recognition Engine", Proc. *EUROSPEECH 2001*: 1691-1694, Denmark, Sept. 2001.

[11] Schröder, M. and Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6, pp. 365-377.

[12] Bechet F., LIA_PHON: un système complet de phonétisation de textes, *Traitement Automatique des Langues – TAL,* v. 42 n. 1, pp. 47-67, 2001.

[13] Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (ed.) *Prosodic Typology -- The Phonology of Intonation and Phrasing*, 2005.

[14] Xavier F. "*Intégration du français au système MaryTTS*", Internship report of the Master 2 ATIAM (Acoustique, Traitement du signal, Informatique, Appliqués à la Musique), Insitut de Recherche en Coordination Acoustique/Musique IRCAM, 2011.

[15] Mironidis T. "*Text-To-Speech systems*", Master Thesis "Computer Systems", Department of Applied Informatics, University of Macedonia, Greece, 2011.

[16] Chollet G., Cochard J.-L., Constantinescu A., Jaboulet C. and Langlais P. "Swiss French PolyPhone and PolyVar: Telephone Speech Databases to Model Inter- and Intra-Speaker Variability", *Linguistic Databases*, ed. J. Nerbonne, Stanford: CSLI (Number 77), 1997. ISBN: 1-57586-092-9.

[17] Kominekand J., Black A.-W. "CMU ARCTIC databases for speech synthesis, " *Techincal Report CMU*-LTI-03-177, Carnegie Mel-lonUniversity, 2003.