9th International Seminar on Speech Production 2011

ISSP

Montreal, Canada, june 20-23 2011

Towards a Practical Silent Speech Interface Based on Vocal Tract Imaging

Bruce Denby[1,2], Jun Cai[1,2], Thomas Hueber[3], Pierre Roussel[2], Gérard Dreyfus[2], Lise Crevier-Buchman[4], Claire Pillot-Loiseau[4], Gérard Chollet[5], Sotiris Manitsaris[1,2], Maureen Stone[6]

[1]Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France

[2]SIGMA Laboratory, ESPCI ParisTech, CNRS-UMR 7084, 10 rue Vauquelin, 75005 Paris, France

[3]GIPSA-Lab, Département Parole & Cognition, CNRS-UMR 5216, 961 rue de la Houille Blanche, Domaine universitaire, BP 46, 38402 Saint Martin d'Hères Cedex France

[4]Laboratoire de Phonétique et Phonologie, CNRS-UMR 7018, 19 rue des Bernardins, 75005 Paris, France

[5]Lab. Traitement et Communication d'Information, CNRS-UMR 5141, Ecole Nationale Superieure des Télécommunications, Telecom-ParisTech, 37-39 rue Dareau, 75014 Paris, France

[6]Vocal Tract Visualization Lab, U. of Maryland Dental School, 650 W. Baltimore St., Baltimore, MD, 21201 USA

denby@ieee.org, Jun.Cai@ieee.org, thomas.hueber@gipsa-lab.grenoble-inp.fr, {pierre.roussel, gerard dreyfus}@espci.fr, lise.crevier@univ-paris3.fr, claire.pillot@univ-paris3.fr, gerard.chollet@telecom-paristech.fr, sotiris.manitsaris@espci.fr, mstone@umaryland.edu

Abstract. *The paper describes advances in the development of an ultrasound silent speech interface for use in silent communications applications or as a speaking aid for persons who have undergone a laryngectomy. It reports some first steps towards making such a device lightweight, portable, interactive, and practical to use. Simple experimental tests of an interactive silent speech interface for everyday applications are described. Possible future improvements including extension to continuous speech and real time operation are discussed.*

1. Introduction

Silent Speech Interfaces, or 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 (Denby et al. 2010). Development of SSIs, using a wide variety of

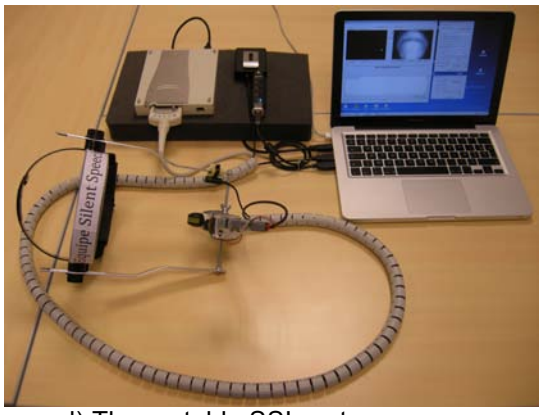
sensor types, remains an active area of research. In Denby et al. (2010) and Hueber et al. (2008a, 2009, 2010), real time ultrasound (US) and video imaging of the vocal tract was shown to be effective for offline, visuo-phonetic continuous speech recognition, in a fixed, "benchtop" SSI. Though interesting as a proof of principle, such a system is impractical for everyday applications since both SSI and user are required to remain immobile. In Florescu et al. (2010), a system employing a professional ultrasound acquisition helmet, with an added camera, was described. 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.

We report here on tests of a new, lightweight SSI, which can be easily carried, is relatively independent of operating environment, and can already be used to perform simple interactive demonstrations. The system is described in section 2. Section 3 presents results from some simple tests of the system on demonstrator applications. Conclusions, more recent developments, and perspectives for the future are discussed in the final section.

a) Lightweight helmet

b) Detail of helmet

c) Taking data with Ultraspeech



d) The portable SSI system

Figure 1. Portable SSI with helmet, IR camera, and miniature US machine.

2. System Description

The SSI system is comprised of a lightweight, adjustable helmet (figure 1a), fitted with a microconvex, 1 inch radius, 128 element US probe for tongue imaging, a VGA, 60 frame per second (fps) CMOS industrial camera for the lips, and a lapel microphone for audio (seen clipped to a vertical strut at the left of figure 1b), located about 5 cm from the speaker's mouth. An infrared (IR) illumination, supplied by IR light emitting diodes (LEDs), and an IR filter mounted on the camera, render video acquisition relatively independent of the ambient lighting conditions.

The US machine used is the Terason t3000™ OEM version (the small white box next to the computer in figures 1c and 1d). The imaging devices are controlled by the Ultraspeech software package (Hueber et al., 2008b), whose multithread programming technique allows simultaneous, synchronous acquisition of the video streams at their maximum frame rates, along with the audio. The Ultraspeech graphical interface allows the user to easily perform visual speech data acquisition. It also provides a method to recalibrate the ultrasound and video images during recording to maintain positioning accuracy (Florescu et al., 2010, Hueber et al., 2008b), by comparing the current image to standard reference images. The entire SSI system can be placed in a small carrying case and operated on battery power, thus enabling practical everyday applications.

3. Tests of the SSI System

Two tests were carried out to test the efficacy of our portable SSI system: first, a telephone dialing test, analyzed offline, to check that the sensors mounted on our home-made, lightweight helmet are still viable in this type of scenario; and second, a simple isolated digit test, performed online, to get a first idea of the practicality of operating such a device in an interactive way.

3.1 Dialing Test

To determine whether the data taken with our new portable system will still be adequate to perform visuo-phonetic speech recognition, a simple telephone dialing application, similar to the one originated in Young et al. (online document), was first developed. The vocabulary chosen contained 3 dialing commands (Call, Phone, and Dial), the digits 0-9, and 34 proper names. The names could be represented either by a first and last name combination, or by last name alone. Four hundred randomly generated dialing instructions were generated from this vocabulary, and uttered once each by a native English speaker using silent speech, that is, without any laryngeal activity.

Data from the US, video, and audio streams recorded by Ultraspeech may be consulted at http://ftp.espci.fr/shadow/Dial_Test2. The audio captured up by the lapel microphone contains only faint clicking noises produced by the mouth during non-vocalized speech, thus corroborating the truly silent signature of the audio signal in this test. For comparison, the audio from a normally vocalized command, using the same audio volume settings, is also included at the mentioned website.

In order to test the robustness of the system to variations in experimental conditions, data were recorded in two sessions of 200 instructions each, one in the morning with window shades open and fluorescent room lights off, and one in the afternoon with window shades open and room lights on; a recalibration (Florescu et al., 2010, Hueber et al., 2008b) was performed between sessions. Visual speech features were extracted using the principal component analysis (PCA) based EigenTongues/EigenLips approach of Hueber (2009). Phone models were trained as cross-word triphone HMMs, and simple "task grammars" used to constrain visual speech decoding. One hundred instructions were chosen at random to form a test set. The achieved word accuracy obtained after training, and then testing the system offline, was 94%. This result gives us confidence that the information extracted by the helmet sensors remains pertinent for visuo-phonetic speech recognition even when the acquisition system used is portable.

3.2 Online Digit Test

As a first test of the online capabilities of our system, an interactive isolated digit recognition demonstrator was also developed. A native English speaker recited the digits 0-9 once each, this time in normal, articulated speech, in order to form a dictionary. The audio track for each utterance – in this case, normal, audible speech – is also stored as a .wav file. The visual features of the resulting images were extracted as the first 30 discrete cosine transform (DCT) coefficients of each image, after first reducing images to 64x64 pixels. The DCT was chosen here to avoid the additional step of performing PCA. Subsequently, when the same speaker repeats a digit, the DCT features of the newly acquired sequence are extracted online, and compared to the stored digit feature sequences using a dynamic time warping (DTW) algorithm, as described in Florescu et al., 2010. After a processing time of a few seconds, the system then plays back the .wav file corresponding to the vocalized utterance from the dictionary that has the closest match. When testing the system, the speaker may pronounce the digit either in normal, vocalized speech, or silently. In the latter case, the system becomes a rudimentary SSI, "transforming" the silent digit into a vocalized one. After a little practice, a 100% digit recognition rate could be achieved in this way, irrespective of whether digits were pronounced silently or aloud. The results suggests that simple, interactive applications of the SSI should be possible, even if, for longer utterances, overall processing times will be proportionately longer.

4. Conclusions, Current Developments, and Perspectives

Our tests show that vocal tract information obtained from ultrasound and video sensors attached to a lightweight, wearable helmet is useful for performing visuo-phonetic speech recognition in online applications. At present, a few seconds are necessary to process short utterances. In future, it should be possible to obtain performance approaching real-time by streamlining the software flow and accelerating the image processing and feature extraction tasks. Software capable of performing real-time acoustic speech recognition has existed for some time. Tests being carried out currently in our laboratories indicate that it should be possible to adapt the Julius real-time speech

recognition system (Lee et al. 2001) to the real-time visuo-phonetic speech recognition task required for our application.

Our tests were, furthermore, carried out within the REVOIX project (ANR 2009), which aims ultimately to produce a portable real time silent speech interface to restore the original voice of individuals who have lost the ability to vocalize speech due to laryngectomy. It should already be possible, using the techniques presented here, to develop domain-specific phrasebook style applications using recordings of patients' voices taken before surgery, in order to at least partially meet the goal of REVOIX. Tests are also being carried out with much larger training corpora, both in French and in English, with the objective of performing continuous-speech visuo-phonetic recognition with our device. In this case, output speech resembling the speaker's original voice can be obtained by feeding the recognized speech text to a text to speech (TTS) system that has been trained on the speaker's voice before his or her operation. Current tests of the open-source OpenMary platform (DFKI, online document) for this task appear promising.

Finally, the current laptop interface, though ideal for data acquisition, is inappropriate for portable operation of the system in real situations. To overcome this limitation, it is possible to keep the laptop closed and use it only as a calculating engine. Access to interactive menus can then be accomplished by running a remote desktop type application on a handheld device over a simple WiFi or 3G wireless Internet connection, as tests with an iPad or iPhone (Apple 2010) in our laboratories have demonstrated. Thus, by using a wearable ultrasound/video acquisition helmet, appropriate real-time image and speech processing software, a portable calculation engine, and a wireless handheld human interface, the possibility of genuinely practical silent speech interface appears rather promising.

References

ANR 2009: French National Research Agency (ANR) contract numbers ANR-09-ETEC-005-01 and 02 REVOIX.

Apple 2010: Apple Computer Corporation, Cupertino, California, USA.

Denby, B. et al. Silent Speech Interfaces. Speech Communication, 52(4):270-287, 2010.

DFKI, http://mary.dfki.de/ OpenMary TTS System.

Florescu, V. et al. Silent *vs* Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface. In *Proceedings of Interspeech 2010, Makuhari Japan*: pages 450-453, September 2010.

Hueber, T. et al. Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips. In *Proceedings of Interspeech 2008, Brisbane, Australia*: pages 2028-2031, September 2008a.

Hueber, T. et al. Acquisition of Ultrasound, Video and Acoustic Speech Data for a Silent-Speech Interface Application. In *Proceedings of ISSP2008, Strasbourg, France*: pages 365-369, December 2008b; see also online at http://www.ultraspeech.com

Hueber, T. Reconstitution de la Parole par Imagerie Ultrasonore et Vidéo de l'Appareil Vocal: vers une Communication Parlée Silencieuse. Doctoral Thesis, Université Pierre et Marie Curie, December 2009 (in French).

Hueber, T. et al. Visuo-Phonetic Speech Decoding Using Multi-Stream and Context-Dependent Models for an Ultrasound-Based Silent Speech Interface. In *Proceedings of Interspeech 2009, Brighton, UK*: pages 640-643, September 2009.

Hueber, T. et al. Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips. Speech Communication, 52(4):288-300, 2010.

Lee, A. Kawahara, T. and Shikano, K. Julius – An Open Source Real-time Large Vocabulary Recognition Engine. In Proceedings of Eurospeech 2001, Denmark, pages 1691-1694, September 2001.

Young, S. Evermann, G. Gales, M. et al., *The HTK Book*, online at http://htk.eng.cam.ac.uk/docs/docs.shtml