TESTS OF AN INTERACTIVE, PHRASEBOOK-STYLE POST-LARYNGECTOMY VOICE-REPLACEMENT SYSTEM

Bruce Denby[a,b], *Jun Cai*[a,b], *Pierre Roussel*[b], *Gérard Dreyfus*[b], *Lise Crevier-Buchman*[c], *Claire Pillot-Loiseau*[c], *Thomas Hueber*[d] & *Gérard Chollet*[e]

[a]Université Pierre et Marie Curie, Paris, France;
[b]SIGMA Laboratory, ESPCI ParisTech, CNRS-UMR 7084, Paris, France;
[c]Laboratoire de Phonétique et Phonologie, CNRS-UMR 7018, Paris, France;
[d]Département Parole & Cognition, GIPSA-Lab, CNRS-UMR 5216, Grenoble, France;
[e]Laboratoire Traitement et Communication de l'Information, Telecom ParisTech, Paris, France
denby@ieee.org; Jun.Cai@ieee.org;
Pierre.Roussel@espci.fr; Gerard.Dreyfus@espci.fr;
lise.buchman@numericable.fr; claire.pillot@univ-paris3.fr;
thomas.hueber@gipsa-lab.grenoble-inp.fr; chollet@telecom-paristech.fr

ABSTRACT

The article presents the results of tests of a portable post-laryngectomy voice replacement system that allows a silently articulating speaker to select and play back short phrases contained in a 60-phrase phrasebook. Such a system could be a useful communication tool for post-laryngectomy patients unable to use tracheo-oesophageal speech. Experiments on two non-pathological speakers and one person having undergone a total laryngectomy in 1998 are presented. Results are promising and provide proof of principle for a more sophisticated system currently being developed.

Keywords: laryngectomy, voice replacement, ultrasound, silent speech interface

1. INTRODUCTION

There has been renewed interest recently [2] in developing post-laryngectomy voice-replacement (VR) technologies as alternatives to oesaphageal (ES) or tracheo-oesophageal (TES) speech, which have a number of practical limitations. Solutions employing a variety of different types of sensors have been proposed, including electromagnetic articulography (EMA) markers, surface electromyography electrodes (sEM), ultrasound (US) tongue imaging, video lip imaging, etc. [2], even if all remain experimental for the moment.

In order to be useful in real life applications, VR systems need to be non-invasive, portable, and as near to real time as possible. Ultrasound imaging furnishes detailed information about the vocal tract without attaching electrodes or markers to the body, and very lightweight US machines are available today. For these reasons, US makes a very good candidate for a VR system.

The Revoix project at the Sigma Laboratory in Paris, France [8], proposes to develop a portable, post-laryngectomy US VR system that will restore the original voice of a patient. The efficacy of *offline* speech recognition and synthesis using US coupled with a lip video camera has already been demonstrated for non-pathological speakers in a laboratory setting [6]. An important first step for Revoix will therefore be to demonstrate that tongue and lip data obtained from actual post-op patients using portable equipment can actually be exploited in an online, interactive scenario. That is the purpose of this article.

We describe tests of a portable, interactive VR device, which were performed upon two non-pathological individuals and one volunteer (henceforth, speaker L) who had undergone a total laryngectomy in 1998 and now speaks with TES. The system allows the user to recall one of 60 pre-recorded phrases from a phrasebook by articulating the phrase silently while fitted with an instrumented helmet. Such a device could serve as a simple tool, for example, for post-operative laryngectomy patients who have not mastered TES. The results we present are promising, and prepare the way for the development of more sophisticated, large vocabulary VR systems using US technology. The speech corpus and data acquisition and analysis procedures used are described in section 2. Results and discussion follow in section 3, while a conclusion and some perspectives are presented in section 4.

2. CORPUS, ACQUISITION & ANALYSIS

The speech corpus (in French) was constructed from a tape recording of everyday conversations involving the post-op volunteer, made before his operation. Sixty phrases of 5 to 10 words were extracted from this tape and transcribed manually to text. The phrases were chosen not on any phonetic equilibrium basis, but rather to assemble, as well as possible given the content and acoustic quality of the recording, a set of intelligible, short sentences of potential everyday use to someone who has lost the ability to speak.

The acquisition helmet, fitted with a 196-element curvilinear 4-8 MHz US transducer, a 60 fps B/W CMOS lip camera illuminated by infrared LEDs, and a small lapel microphone, is illustrated in figure 1. The US machine [9] and portable computer fit in a small carrying case, making the system portable.

Figure 1: Acquisition helmet worn by post-laryngectomy volunteer; ultrasound transducer beneath the chin; 60 fps lip camera with IR illumination; lapel microphone (clipped to aluminum strut). Tracheostoma is visible to right of transducer.

Each speaker first made a training pass by repeating each phrase silently once while wearing the helmet, with the US, video, and audio streams being logged to the computer via the Ultraspeech software package [10]. Since silent speech is used here, the audio stream is not really necessary; however, interested readers may consult samples of the audio recorded by the microphone by at [1]. The training pass is followed by a ~1h pause during which the speaker removes the helmet and relaxes while the feature extraction is performed on the recorded data. In this step, a discrete cosine transform (DCT) is applied to the US and lip images, and the first 30 DCT coefficients of each retained. The first and second derivatives of these coefficients are also included, so that each tongue and lip image is represented by a 90-element feature vector in a training database. After the pause, the helmet is returned to the speaker's head and repositioned using a calibration procedure described in [4].

In the subsequent testing pass, the speaker repeats each phrase again silently once. After each phrase, an online software application calculates the feature vectors of the images just recorded by Ultraspeech, and uses a dynamic time warping algorithm (DTW) to determine the training set phrase whose sequence of stored feature vectors best matches that of the phrase just pronounced [3]. The pre-recorded audio clip corresponding to this phrase (pre-op voice of speaker L), after a delay of a few seconds, then plays back over the loudspeaker of the PC. If the found phrase corresponds to the phrase that was actually pronounced, a correct response is counted. The percentage of phrases correctly recognized gives an indication of the performance of the VR system. The DTW can also be configured to use only the tongue or only the lips, in order the test the efficacy of each articulator taken independently. For one of the speakers, additional test passes were also performed at later dates, in order to assess the time stability of the system.

3. RESULTS AND DISCUSSION

The performance scores for the three speakers are presented in Table 1. The post-laryngectomy volunteer is labeled L in the table, the other two speakers, S1 and S2.

Table 1: Number of correct sentences out of 60 and percentage score for the post-op speaker, L, and other two speakers, S1, S2.

Speaker	No. correct	% correct
S1	57/60	95%
S2	60/60	100%
L	50/60	83%

The performance for speakers S1 and S2 is excellent, while that of L is somewhat lower. Although speakers S1 and S2 were already familiar with the acquisition procedure (essentially, clicking an onscreen button, speaking, and clicking again), speaker L was inexperienced and sometimes clicked too early or too late, which causes problems for a simple algorithm such as a DTW. The score of speaker L could thus probably be improved with practice.

The fact that speaker L was able to achieve good performance with our system is encouraging, and was far from obvious before the experiment. Indeed the vocal tract of a post-laryngectomy patient differs substantially from that of a normal tract. For example, the hyoid bone shadow was absent in the tongue images of speaker L, since the hyoid was removed during the operation. The tongue images of L were however of good quality, and the DCT was ostensibly able to extract discriminating information from them. The beard and moustache of speaker L furthermore did not appear to cause problems with US acoustic contact or with the feature extraction of the lip images. Finally, it was verified that the US probe would not interfere with the tracheostoma of speaker L, as is seen in figure 1.

In interpreting the results, it is important to verify that the discriminatory capability we are observing is not based on some artefact such as the length of the sentence (albeit this is technically compensated for in a DTW). An examination of phrases that the system misclassifies can offer insight on this point. For a DTW with an articulatory distance measure, and with traditional assumptions about system noise, the most probable confusions should take place between the phrases that are closest in articulatory space. Our corpus contains a handful of phrases with very similar phonetic transcriptions, for example:

- "I nous ont jamais répondu"
  ii nn ou zz on jj aa mm ai rr ei pp on dd uu
- "Il m'a jamais répondu"
  ii ll mm aa jj aa mm ai rr ei pp on dd uu;

where the ends of the two sentences are identical. The bulk of the errors observed (for table 1 and table 2, described below) indeed corresponded either to confusions of these two phrases or to other, similar cases. Had the DTW been keying off unrelated information, a random distribution of system confusions would rather have been expected.

Table 2 presents the results of further tests performed on speaker S2 at later dates. The individual contributions of the tongue and lips to the decisions taken by the algorithm are also separated out here, in order to check for possible problems in one branch or the other.

Table 2: Number of correct sentences out of 60 and percentage score for speaker S2 after 0, 2 and 3 weeks, also showing contribution of lips and tongue separately.

Speaker S2		No. correct	% correct
0 wk	both	60/60	100%
	tongue	59/60	98%
	lips	60/60	100%
2 wk	both	49/60	82%
	tongue	50/60	83%
	lips	35/60	58%
3 wk	both	56/60	93%
	tongue	55/60	92%
	lips	40/60	67%

The table shows that when the testing is performed one hour after the training, very nearly 100% performance is achieved, but that in new sessions performed at later times, the performance degrades. The degradation is modest in the tongue channel, 83-92% instead of 98%, but quite dramatic for the lips, 58-67% rather than 100%. The reason for these degradations is not yet understood; however, it is suspected that the system is quite sensitive to precise recalibration between sessions.

The results at zero weeks could lead one to ask if we could eliminate the US transducer and just use the camera, or perhaps even some simpler device, to perform the same 60-phrase task. While that is open to investigation, it will clearly not be an option for future VR systems like the one planned in Revoix, which will allow the speaker to pronounce arbitrary phrases at will. Indeed, earlier research [5] on a US + camera system has demonstrated that for visuo-acoustic continuous speech recognition using Hidden Markov Models (HMM), the tongue and lip channels are complementary, and it is the tongue that conveys the major portion of the articulatory information. In any event, for our system, the lip camera channel proved much less stable over time than the US channel.

4. CONCLUSIONS AND PERSPECTIVES

We have demonstrated that a speaker who has undergone a total laryngectomy can make fruitful

use of an interactive, phrasebook-style VR system based on US images of the tongue and a lip video camera. Using a simple online algorithm that extracts and compares articulatory information from the acquired images, the correct phrase can be selected and played back from a book of 60 short phrases more than 80% of the time in a few seconds. Results present some variability over time, particularly in the lip channel, for reasons that are not yet understood. No particular problems appear to be posed by the vocal tract configuration or tracheostoma of the laryngectomized person. The system presented could be useful even in its present form for post-laryngectomy patients who have not learned to use TES. It has the further advantage of reproducing the original voice of the patient, which could be an attractive feature for the user.

The more sophisticated VR being developed in Revoix will perform visuo-acoustic speech recognition using HMMs coupled with a statistical language model in order to allow the speaker to produce any phrase. Speech synthesis will be carried out from the decoded text using a Text To Speech system (TTS) that has been trained on the original voice of the speaker. A DTW algorithm was used in the present demonstrator because of its rapidity of execution for an online application. In the final system, real time performance should be possible with HMMs as well by using the Julius speech recognition system [7]. We will of course want to continue to improve the ergonomics of our acquisition helmet. We plan to test speaker-specific thermoformed helmets in upcoming experiments, for example. The results presented have provided us with some clues for further improving the time stability of our system. Finally, we will extend our studies to multiple speakers, to ensure the universal applicability of the techniques proposed.

5. ACKNOWLEDGMENTS

This work was supported by the Agence Nationale de la Recherche under contracts number ANR-09-ETEC-005-01&2 REVOIX.

6. REFERENCES

[1] Audio samples *http://ftp.espci.fr/shadow/SSI_Test/*

[2] Denby, B. et al. 2010. Silent speech interfaces. *Speech Comm*. 52(4), 270-287.

[3] Florescu, V., et al. 2010. Silent vs vocalized articulation for a portable ultrasound-based silent speech interface. *Proc. Interspeech 2010* Japan, 450-453.

[4] Hueber, T., et al. 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. ISSP2008* Strasbourg, France, 365-369.

[5] Hueber, T., et al. 2009. Visuo-phonetic decoding using multi-stream and context-dependent models for an ultrasound-based silent speech interface. *Proc. Interspeech 2009* UK, 640-643.

[6] Hueber, T., et al. 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Comm*. 52(4), 288-300.

[7] Lee, A., Kawahara, T., Shikano, K. 2001. Julius – An open source real-time large vocabulary recognition engine. *Proc. Eurospeech 200*1 Denmark, 1691-1694.

[8] Revoix project (Sigma Laboratory in Paris, France) *http://www.neurones.espci.fr/Revoix/*

[9] Terason Corporation, Burlington, Massachusetts, USA *http://www.terason.com/*

[10] Ultraspeech software *http://www.ultraspeech.com/*