Vocal Tract Imaging System for Post-Laryngectomy Voice Replacement

Jun Cai(M'04)[1,2], Thomas Hueber[3], Sotiris Manitsaris[1,2], Pierre Roussel[2], Lise Crevier-Buchman[4], Maureen Stone[5], Claire Pillot-Loiseau[4], Gérard Chollet[6], Gérard Dreyfus(F'12)[2], Bruce Denby(SM'99)[1,2]

[1]Université Pierre et Marie Curie, Paris, France
[2]SIGMA Laboratory, ESPCI ParisTech, CNRS-UMR 7084, Paris, France
[3]GIPSA-lab, UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France
[4]Laboratoire de Phonétique et Phonologie, CNRS-UMR 7018, Paris, France
[5]Vocal Tract Visualization Lab, University of Maryland Dental School, Baltimore, MD, USA
[6]Laboratoire Traitement et Communication de l'Information, CNRS-UMR 5141, Paris, France
denby@ieee.org

***Abstract*—The article describes a system that uses real time measurements of the vocal tract to drive a voice-replacement system for post-laryngectomy patients. Based on a thermoformed acquisition helmet, miniature ultrasound machine, and video camera, and incorporating Hidden Markov Model speech recognition, the device has been tested on three speakers, one of whom has undergone a total laryngectomy. Results show that the device obtains exploitable recognition rates, and that performances on normal and post-laryngectomy speakers are nearly identical. The technique can also enable voice communication for normal speakers in situations where silence must be maintained.**

Keywords—vocal tract measurement; silent speech interface; voice replacement; handicapped speech; laryngectomy

I. INTRODUCTION

Each year, thousands of persons worldwide fall victim to a cancer of the larynx or a neurological problem leading to loss of voice, constituting a serious social handicap. Unfortunately, the traditional solutions, such as the electrolarynx or tracheo-esophageal speech, TES, provide a replacement voice with a rather unnatural sound, which can be awkward, both for the speaker and for those nearby. Consequently, the investigation of new sensor-based systems for handicapped voice replacement has remained an active area of research, with a wide range of candidate technologies being proposed in the literature [1-8]. Many of the proposed systems can also double as so-called "silent speech interfaces" [1], which allow handicapped or non-handicapped speakers to perform speech communication in situations where silence is required, simply by articulating normally, but without any vocal chord activity.

In [3-5], results were presented on such a system for native English speakers, making use of ultrasound and video imaging of the tongue and lips to drive a "visual speech" recognition engine. In the present article, we extend this technique to the French language, and, for the very first time, evaluate the approach on a post-laryngectomy patient. Since laryngectomy surgery leaves the tongue, lips, teeth, etc., untouched, such patients should not, a priori, be at a disadvantage in using the proposed technique, as compared to normal speakers. Our results confirm that this is indeed the case, and indicate that the device, coupled with a voice synthesizer, can indeed form an integral part of a lightweight, practical voice-replacement system for post-laryngectomy speakers.

The data acquisition system used in the system is outlined in section II, while the speech corpora (sets of training and validation utterances to be pronounced by each speaker) developed are described in section III. Experimental results appear in section IV, followed by discussion in section V, and conclusions and future perspectives in section VI.

Figure 1. Speaker in sound cabin with form-fitted plastic acquisition helmet. Ultrasound probe is on white cable entering from below, lip camera is small blue box in front of mouth. Terason T3000 ultrasound machine is on small tray beneath portable PC running Ultraspeech.

II. DATA ACQUISITION AND PREPROCESSING

Each speaker is outfitted with a lightweight plastic helmet (approximately 400 g) holding a Terason 8MC4, 4-8 MHz,

140°, 128-element micro-convex ultrasound transducer under the chin for tongue imaging, a front-mounted 640×480 pixel black and white CMOS industrial camera for the lips, and a lapel microphone for recording sound (figure 1). In contrast to the acquisition platform developed in [3-5], these helmets are form-fitted individually for each speaker using a computer-assisted prototyping system, in order to create a stable, head-referenced platform for the ultrasound probe and camera with a minimum of adjustable, moving parts. Cameras are equipped with an infrared (IR) filter and ring of IR Light Emitting Diodes (LED) in order to render lip imaging independent of ambient lighting conditions. The ultrasound machine, a Terason t3000™, is about the size and weight of a small paperback book. Standard ultrasound gel is applied regularly at the interface between transducer and skin in order to ensure good acoustic coupling.

Acquisition is controlled by a stand-alone dedicated graphical interface, Ultraspeech (www.ultraspeech.com, [8]), which uses a multithread programming technique to allow synchronous acquisition of the two image streams at their respective maximum frame rates, along with an audio signal. At an ultrasound focal distance of 7 cm, appropriate for tongue visualization, the system simultaneously and synchronously records the ultrasound stream at 60 fps (image resolution of 320×240 pixels); the video stream at 60 fps (image resolution of 640×480 pixels); and the audio signal (16 KHz, 16 bits). The program runs on an ordinary laptop PC so that the entire system can be transported in a small carrying case if required. Recording a large amount of multimodal speech data for training recognition models requires multiple sessions spaced in time, necessitating inter-session re-calibration to maintain positions of sensors and readjustment between sessions. Ultraspeech allows the user to interactively re-calibrate ultrasound and video images periodically by comparing current images to pre-recorded reference images, as described in [9].

The acquired images are preprocessed by selecting a Region Of Interest (ROI) containing tongue or lips, to which a Discrete Cosine Transform (DCT) is applied, after resizing to 64×64 pixels. The first 30 DCT coefficients, with their first and second derivatives, form a 90 (tongue) + 90 (lips) = 180-component feature vector for each image. These vectors are then passed to the Hidden Markov Model, or HMM-based, training/recognition procedure described in section IV.

III. Speech Corpora and Language Model

Effective HMM training/recognition requires a rather large training database. French was chosen as the target language in order to allow direct comparison with our post-laryngectomy collaborator, who was a French-speaker. Unfortunately, standardized corpora for creating a database in the French language are less accessible than those available for English. Consequently, purpose-built corpora had to be developed specifically for our experiments, using the French Polyvar text corpus [10] as a basis.

A. The Ω Corpus of Short Sentences

Polyvar contains 44,762 sentences from orthographic transcriptions of a Swiss French telephone speech corpus. To ensure sentences used were neither too short, nor exceeded our system's acquisition buffer of 8 seconds, a subset of these, Ω, was first extracted, containing 11,669 sentences of 60 to 110 characters.

B. Creating and Recording Training Corpora

The basic subunits used in continuous speech recognition are usually context-dependent triphones, chains of three phonemes that model each phoneme in a variety of contexts depending on the phonemes immediately preceding and following it, i.e., on its "left and right context". In this way, what are referred to as co-articulation effects are modeled. Cross-word triphone HMMs, which allow triphones to cross word boundaries, will be the basic building blocks of our visual speech recognition system.

As recording ultrasound speech data is an involved and time-consuming activity, it is desirable to keep the size of the training corpus as small as possible, while at the same time assuring that all phonemes are adequately covered. To this end, a "greedy" algorithm [11] was used to extract sentences one by one from the set Ω, such that each new sentence introduced as many new triphones as possible. To enable the operation of the greedy algorithm, the sentences in Ω were transcribed into phonemes using the French Text-To-Phoneme converter lia_phon [12] that automatically generates French phonetic transcriptions.

The greedy algorithm was stopped once 3000 sentences were extracted from Ω. These 3000 sentences contain 14,638 triphones out of the total of 15,119 in Ω, that is, more than 96% of all triphones contained in Ω. As the triphone content of the sentences picked by the greedy algorithms increases with each successive iteration, a random shuffle was applied to these 3000 sentences in order to assure a uniform distribution of triphones through the training set as it is presented to the training algorithm.

Two male native French speakers, PRR and JCD, and a female native French speaker, LCB, recorded French visual speech data, i.e., ultrasound images of the tongue and visual images of the lips, using the described training corpus. Speaker JCD, who had undergone a total laryngectomy several years before these tests, today uses TES, and breathes through a tracheostoma. As expected, tongue images for JCD were found to be similar to those of the other speakers, except for the absence of the shadow of the hyoid bone, which had been removed in his surgery. The helmet was furthermore well tolerated by this speaker (e.g., no interference with the tracheostoma), and acquisition protocol posed no particular problems.

In order to simulate the use-case of a post-laryngectomy speaker, all speakers were instructed to completely suppress glottal (vocal chord) activity while recording the texts, that is, to speak silently, moving only the tongue and lips. Whispering was also excluded, as whispered articulation is known to be different from vocalized articulation. For speaker JCD, who has no glottis due to his operation, this amounted to simply not installing his one-way TES valve. Articulation was performed at a normal pace, in a clear, but non-exaggerated way, and as consistently as possible. Although an audio track was also

recorded, it is used only for control purposes during this "silent speech" procedure, to verify that glottal activity was indeed kept to an absolute minimum.

For speaker PRR, all 3000 sentences were recorded, while for speakers JCD and LCB, in order to further reduce the acquisition time, a smaller subset of 2000 sentences was used. For each speaker, 6% of the visual speech data was selected in a random way from the training set and retained as test data, with the remainder of the data being used for model selection and HMM training. The resulting numbers of training and test sentences for the three speakers are listed in Table I. The same set of test sentences was used for speakers LCB and JCD.

C. Language Model

In order to improve the recognition performance, lexical information based on probabilities of different sequences of words can be added via a statistical Language Model, or LM. For example, in English, the combination "the door" is a priori more probable than "them door". A LM based on all possible pairs of words is referred to as a 'bigram" LM. Thus, using the French short sentence set Ω, a dedicated French bigram LM called here the "Polyvar bigram" was built, for each speaker, for use in performing the test evaluation. To avoid biasing the system towards the training set used in making the LM, the test sentences were first excluded from Ω, and the remaining sentences then used to build the LM. A typical vocabulary size for the resulting bigram LMs is 16,852 words. In order to evaluate the intrinsic quality of the HMM recognition without the LM, as well as illustrate the impact of the LM for our application, we also tested a simple "word-loop" bigram LM with all word pair probabilities set equal, as proposed in [9].

TABLE I. GENDER AND NUMBER OF TRAINING AND TEST SENTENCES FOR THE THREE SPEAKERS.

Speaker Name	Speaker Gender	No. of Training Sentences	No. of Test Sentences
PRR	Male	2820	180
LCB	Female	1880	120
JCD[a]	Male	1880	120

[a]post-laryngectomy speaker

IV. EXPERIMENTAL RESULTS

Using the 180-component visual speech DCT feature vectors from the training corpus as input, the HTK 3.4 HMM toolkit [13] was used to train context-dependent, cross-word triphone HMMs, of 3-state, left-to-right topology. Each state, in this case, refers to a phoneme (or a special label, as below). The probability density functions describing the various states are estimated using Gaussian Mixture Models (GMM), during the training phase. Based on an empirical study, each HMM state of a phoneme was modeled by an 8-gaussian Gaussian Mixture Model (GMM), while the special HMM states SILENCE and SHORT_PAUSE, used for representing pauses of different lengths, required 32-gaussian GMMs. In the decoding (i.e., test) stage, for each speaker, 20 test sentences were used as a development set to tune the so-called "model insertion penalty" (which counteracts a known tendency of the system to insert spurious short words in an effort to improve the score) and LM factor, until the numbers of errors due to word insertion and word deletion were approximately balanced, as is standard procedure. Word-level and phone-level recognition results for the three speakers, for the Polyvar and Word Loop bigram LMs, are presented in Table II.

TABLE II. WORD AND PHONE LEVEL RECOGNITION ACCURACY IN % FOR THE 3 SPEAKERS AND 2 LMS, WITH 95% C.L. INTERVALS.

Speaker	Recognition Accuracy in %			
	Word Level		*Phone Level*	
	Word Loop Bigram	Polyvar Bigram	Word Loop Bigram	Polyvar Bigram
PRR	10.1±2.4	**64.0±3.8**	62.1±2.0	**77.9±1.7**
LCB	12.4±3.2	**65.7±4.5**	65.0±2.4	**78.7±2.1**
JCD[a]	29.0±4.3	**65.4±4.5**	71.3±2.3	**74.6±2.2**

[a]post-laryngectomy speaker

The table shows that good recognition scores, both at word and phone level, were obtained by all three speakers, and that the scores are relatively uniform across the three speakers. In particular, the scores of the post-laryngectomy speaker were no worse, and in some cases, were better, than those of the two normal speakers. This is a very encouraging result, since it confirms the expectation that a laryngectomized speaker will not be at a disadvantage using the method. Indeed, as articulation is, with very few exceptions, the province of the oral and pharyngeal cavities, it is almost completely unaffected by the removal of the larynx. It furthermore appears that, a priori, increasing the number of training sentences from 2000 (LCB, JCD) to 3000 (PRR) did not lead to better recognition scores. Finally, it is clear from the table that, at word-level, the system relies heavily on the LM to find the correct sequences (about 17% recognition on average without LM versus 65% when using the LM); however, it is instructive to examine some of the output sentences in detail.

To get a more intuitive feel for the kind of performance obtained, two sentences as recognized for speakers LCB and JCD, using the Polyvar bigram, are presented in Table III, along with the original text of the sentences.

TABLE III. TWO SENTENCES AS RECOGNIZED FOR SPEAKERS LCB AND JCD, ALONG WITH ORIGINAL SENTENCE TEXT.

Original	**Et que les dates soient choisies en fonction des besoins de l'opéra non de ses besoins à lui.**
LCB	L'école est à jouer choisir fonction des besoins de l'opéra de son de des besoins lui.
JCD[a]	Et que les dates soient choisies en fonction des besoins de l'opéra non de ses besoins à lui de.
Original	**Les machines et outils sont sculptés dans du bois du bouchon de la feuille de plomb.**
LCB	Les machines est aussi sont sculptés dans du bois du bouchon de la feuille de plomb.
JCD[a]	Les machines et outils sont sculptés dans du bois du bouchon de la fin de plomb.

[a]post-laryngectomy speaker

V. DISCUSSION

In tests of a similar system on a single native English speaker [3-5], the word and phone level recognition results were on average about 12 percentage points higher than those obtained here. Performance differences between French and English have also been reported in the literature for acoustic speech recognition (see for example [14]), and are predominantly attributed to differences of pronunciation and grammar. The quality of the "home-made" speech corpora and LM used for our tests in French are undoubtedly also somewhat to blame for the lower performance.

Nonetheless, examination of the results in table III shows that although the words recognized are sometimes incorrect (see Table I), they often are similar to the correct ones, and would be pronounced in a similar, or sometimes identical, way. Indeed, when recognized sentences are read aloud or synthesized, errors are less apparent, and the meaning of the phrase is at least partially preserved. The conclusion is that the recognition performance obtained here should already be exploitable in real-life systems.

It is notable, furthermore, that the recognition scores of the normal speakers (LCB, PRR) are comparable to that of JCD, the post-laryngectomy speaker. This is consistent with the expectation that oral structures other than the larynx are not affected by the laryngectomy surgery, and that patients of this sort are therefore very good candidates for such a system. In some cases of very extensive surgery, the tongue can be affected as well, which could have adverse affects, however it seems clear from the results presented that typical post-operation patients are just as able to produce consistent datasets as normal speakers, particularly when one considers that the data collection for the training corpus was quite extensive, with each speaker repeating 2000-3000 sentences without vocal chord activity, in sessions taking place over a period of weeks or months.

VI. CONCLUSIONS AND PERSPECTIVES

It has been shown that a lightweight, portable, ultrasound-based vocal tract imaging system can be used effectively for visual speech recognition for post-laryngectomy speakers, thus forming an integral part of a voice-replacement system. As mentioned earlier, the system could also be used as a "silent speech interface" for handicapped or "normal" speakers, who operate the device by articulating without vocal chord activation. The speech recognition performance was also found to be significantly enhanced through the use of a statistical language model. Future work will focus on:

- Improving the recognition performance: We would like to achieve parity with earlier tests in English, and, hopefully, even go beyond that performance. More sophisticated experimental and methodological techniques, including addressing the particularities of the French articulatory paradigm, can likely provide a higher recognition rate.
- Implementing the voice restoration step: The objective of the project is to restore an intelligible speech signal from articulatory movements captured by ultrasound and video sensors. Real-time tests of the recognition step have been carried out in experiments coupling the visual speech recognition system to a Text-To-Speech system (TTS) that has been trained on the original voice of the user [15]. The possibility of having a synthesized voice created from the user's original voice is potentially a very attractive feature for post-laryngectomy speakers, who may today be using an electro-larynx or TES. To date, the synthesis system used was implemented on a remote server; in future tests, we will want to integrate the system directly into the voice-replacement device itself. A disadvantage of the TTS approach method, on the other hand, is that any recognition errors will automatically be reflected in the synthesized speech. It will therefore also be interesting to investigate a complementary approach, combining HMM-based recognition with HMM-based synthesis, in a common statistical framework, as described in [16]. In this method, the speech synthesizer combines recognition results with articulatory information, and could thus possibly provide a more intelligible speech output.
- Improving the system ergonomy: The acquisition helmet will have to be wearable in a practical system. Although lightweight, the current helmet is not genuinely suitable, from aesthetic and ergonomic viewpoints, for application in everyday use cases. Future models will incorporate more highly miniaturized ultrasound and video sensors in a more aesthetic and unobtrusive assembly.

ACKNOWLEDGMENTS

This work was supported by the French National Research Agency (ANR) under contract numbers ANR-09-ETEC-005-01 and ANR-09-ETEC-005-02 REVOIX, and by the French Ministry of Research.

REFERENCES

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert and J.S. Brumberg, "Silent Speech Interfaces", in Speech Communication, Elsevier, Eds., vol. 52, no. 4, pp. 270-287, April 2010.

[2] R. Hofe, S.R. Ell, M.J. Fagan, J.M. Gilbert, P.D. Green, R.K. Moore, S.I. Rybchenko, "Speech Synthesis Parameter Generation for the Assistive Silent Speech Interface MVOCA", in Proc. of Interspeech, pp. 3009-3012, Florence, Italy, 2011.

[3] T. Hueber, E.L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, M. Stone, "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips", in Speech Communication, Elsevier, Eds., vol. 52, no. 4, pp. 288-300, Apr 2010.

[4] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface," in Proceedings of Interspeech, Brighton, UK, pp. 640-643, 2009.

[5] J. Cai, B. Denby, P. Roussel, G. Dreyfus, L. Crevier-Buchman, "Recognition and Real Time Performances of a Lightweight Ultrasound Based Silent Speech Interface Employing a Language Model", in Proc. Interspeech, pp. 1005-1008, Florence, Italy, 2011.

[6] M. Wand, M. Janke, T. Schultz, “Investigations on Speaking Mode Discrepancies in EMG-Based Speech Recognition”, in Proc. Interspeech, pp. 601-604, Florence, Italy, 2011.

[7] C. Herff, M. Janke, M. Wand, T. Schultz, “Impact of Different Feedback Mechanisms in EMG-Based Speech Recognition”, in Proc. Interspeech, pp. 2213-2216, Florence, Italy, 2011.

[8] T. Hueber, G. Chollet, B. Denby, M. Stone, “Acquisition of Ultrasound, Video and Acoustic Speech Data for a Silent-speech Interface Application”, in Proc. International Seminar on Speech Production, pp. 365-369, Strasbourg, France, Dec. 2008.

[9] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, M. Stone, “Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips”, in Proc. Interspeech, pp. 2028-2031, Brisbane, 2008.

[10] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, Ph. Langlais, “Swiss French PolyPhone and PolyVar: Telephone Speech Databases to Model Inter- and Intra-Speaker Variability”, 1996. doi: 10.1.1.49.8609, accessed on Mar. 20, 2012.

[11] H. François, O. Boëffard, “The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database”, in Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Vol. 5, pp. 1420-1426, 2002.

[12] F. Bechet, “LIA-PHON: Un Système Complet de Phonétisation de Textes”, Traitement Automatique des Langues”, vol. 42, no 1, pp. 47-67, 2001.

[13] S. Young, G. Evermann, M. Gales, S. Young, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, “The HTK Book”, Online: http://htk.eng.cam.ac.uk/docs/docs.shtml, accessed on 15 Apr. 2010.

[14] J.-L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, “Large Vocabulary Speech Recognition in English and French”, in Proc. IEEE SPS Workshop on Automatic Speech Recognition, Snowbird, Utah, December 1993.

[15] S. Manitsaris, B. Denby, F. Xavier, J. Cai, M. Stone, P. Roussel, G. Dreyfus, “An Open Source Speech Synthesis Module for a Visual-Speech Recognition System”, in Proceedings of Acoustics 2012, Nantes, France, April 2012.

[16] T. Hueber, G. Bailly, B. Denby, “Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface”, Proceedings of Interspeech, Portland, USA, 2012.