Silent *vs* Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface

Victoria-M. Florescu[1], *Lise Crevier-Buchman*[2], *Bruce Denby*[3,1], *Thomas Hueber*[4], *Antonia Colazo-Simon*[2], *Claire Pillot-Loiseau*[2], *Pierre Roussel*[1], *Cédric Gendrot*[2], *Sophie Quattrocchi*[2]

[1] SIGMA Laboratory, ESPCI ParisTech, CNRS-UMR 7084, Paris, France
[2] Laboratoire de Phonétique et Phonologie, CNRS-UMR 7018, Paris, France
[3] Université Pierre et Marie Curie, Paris, France
[4] GIPSA-Lab, Département Parole & Cognition, CNRS-UMR 5216, Grenoble, France

mihaela.florescu@espci.fr

Abstract

Silent Speech Interfaces have been proposed for communication in silent conditions or as a new means of restoring the voice of persons who have undergone a laryngectomy. To operate such a device, the user must articulate silently. Isolated word recognition tests performed with fixed and portable ultrasound based silent speech interface equipment show that systems trained on vocalized speech exhibit reduced performance when tested on silent articulation, but that training with silently articulated speech allows to recover much of this loss.

Index Terms: silent speech interface, ultrasound, articulation

1. Introduction

The idea of a Silent Speech Interface, or SSI, has received considerable of attention recently in the speech research community [1]. These devices, which perform speech recognition on signals obtained from sensors applied to the vocal tract, are intended to allow their user to communicate anywhere in silence (for example, a "silent cellphone"), or to provide an alternative to tracheo-oesophageal speech for persons who have lost the ability to speak due to laryngectomy. SSIs are still in the experimental phase, and if they are to become genuinely useful, additional breakthroughs will be necessary. For example, the devices will ultimately need to be as non-invasive as possible, and training data for the recognition algorithms reasonably easy to obtain. In [2-4], an SSI based on real time imaging of the tongue and lips using a portable ultrasound machine and a video camera is proposed. Although a promising continuous speech phone recognition rate of 70% on an English corpus was reported in [3], [4], two critical experimental issues remain to be addressed:

- The speaker's head remained immobilized in an acquisition system fixed to a table. Clearly, a practicable SSI will have to be portable.
- The training and test data used in the recognition tests contained standard, vocalized speech. It is nonetheless reasonable to imagine – for example due to the lack of audio feedback – that the articulation process may be different in silent speech. As SSIs are based exclusively on sensor information obtained from the articulators, one may question whether such a training scenario will be viable in a real world situation.

In the present article, we describe isolated word speech recognition tests in which both vocalized and silent speech are experimented in the train and test phases. In addition, both a fixed acquisition system, and a new, portable system, were tested. The results obtained show that a portable acquisition system appears to be feasible, and that significant differences, as reflected in the recognition scores obtained, do indeed exist between articulation in normal vocalized speech and in silent speech. The mechanical and software aspects of the acquisition systems are detailed in the following section, while the speech corpus chosen and databases recorded are described in section 3. The recognition algorithm is outlined in section 4, with the results presented in section 5. Conclusions and some future perspectives appear in the final section.

2. Acquisition systems

A number of different types of ultrasound/video acquisition systems have been described in the literature [5], [6]. Here, a fixed and a portable acquisition system, both using the same ultrasound transducer and CCD video camera, were compared, in order to ascertain if the two sensor mounting approaches exhibited significant differences in performance. The sensors common to both systems are:

- a Terason T3000 OEM[1] ultrasound system with a 8MC4 140° micro-convex transducer, providing tongue images;
- an industrial CCD video camera (for frontal lip images), fitted with a LED ring light.

2.1. Fixed acquisition system

In the fixed system, the transducer and video camera, as well as a microphone, are fixed to a table, as shown in Fig. 1. The subject's forehead presses against an opthalmologists's band, while the ultrasound probe is fixed behind the chin using an articulated arm. The headband, along with cushioned guides on the sides of the head, assure that the head remains fixed and at a constant distance from the video camera and microphone. Using a repositioning technique that will be described in section 2.3, the subject can easily disengage from time to time during multiple sessions. The design of the system is intended to fix the ultrasound probe with respect to the palate, however since the head is free, some relative movement can occur. The fact that the system is attached to a table, clearly, prevents its being used in arbitrary locations.

[1] Terason Ultrasound, http://www.terason.com/products/t3000.asp

Copyright © 2010 ISCA

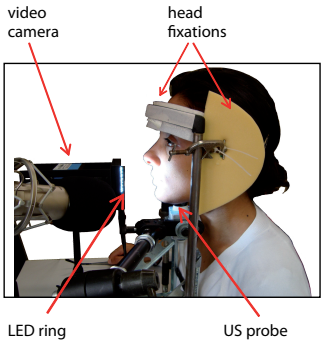

Figure 1: Fixed acquisition system

2.2. Portable acquisition system

The portable data acquisition system makes use of an ultrasound probe stabilization helmet commercialized by Articulate Instruments Ltd. [7]. This device ensures that the ultrasound probe moves rigidly with respect to pressure points on the frontal, occipital, and zygomatic bones during head movement, thus keeping its orientation in regard to the palate fixed. In order to also acquire frontal lip video images, the CCD camera and LED ring were placed on a small horizontal platform attached to the anterior part of the helmet by a pair of adjustable slides, as shown in Fig. 2.

The system maintains the ultrasound probe fixed with respect to the palate, and video camera at a fixed distance from the lips, without restricting head movement. Although somewhat ungainly – and according to our subjects, with its added camera and lighting system, too heavy for long sessions – this portable system can be thought of as an intermediate step on the way to a truly practicable, wearable apparatus.

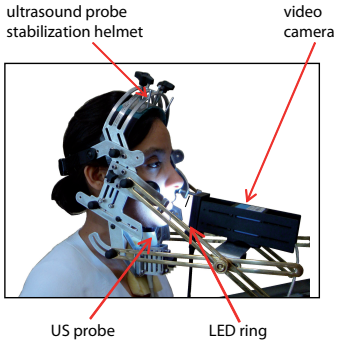

Figure 2: Portable acquisition system.

2.3. Ultraspeech software

In both the fixed and portable systems, the imaging devices (as well as the sound system) are automated by a stand-alone, dedicated software program entitled Ultraspeech [8]. Ultraspeech is a "user friendly" graphical application that allows the synchronous recording of the two image streams and the audio signal at their maximum respective frame rates. The visual and audio streams are processed in parallel using multithreading programming techniques. Streams share a common timer so that each frame and each audio buffer can be tagged with a time value during recording. Any initial asynchrony between streams is captured during the acquisition; synchrony is restored automatically in a post-processing stage. The entire recording procedure is fully automatic in a push-button fashion, with no *a posteriori* human checking required.

After each acquisition, data are directly available as series of bitmaps for the image streams, and .wav files for the audio. With an ultrasound focal distance of 7 cm, which is appropriate for tongue visualization, the system is used to record simultaneously, and synchronously: the ultrasound stream at 60 fps (with an image resolution of 320x240 pixels); the video stream at 60 fps (640x480 pixels); and the acoustic signal (16 bits, 16 kHz).

Ultraspeech provides also an interactive inter-session re-calibration mechanism that allows recording of large audiovisual speech databases in multiple acquisition sessions. The procedure shown in figure 3 is based on real-time averaging of a live image with a target reference image. During this interactive re-calibration procedure, the subject adjusts the position of his/her head in order to fit to the target reference position. A similar procedure is used for ultrasound, where the live tongue image is super-imposed on a target reference.

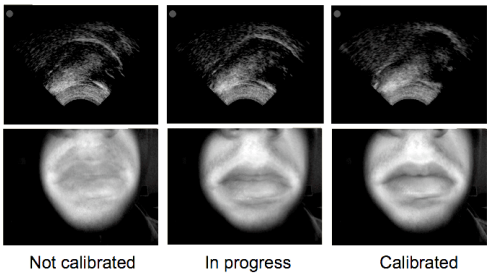

Figure 3: Interactive speaker inter-session recalibration mechanism at different stages of the procedure.

3. Database acquisition

In order to be as close as possible to a useful daily life application, we selected a list of the 50 French words (listed in Table 1) among those that are the most frequent according to the National Education Corpus[2]. This selection was done so that the corpus would contain at least one realization of each French phoneme.

Table 1. *The 50-word corpus*

Words				
le	son	pas	nous	ou
de	que	vous	comme	sans
un	se	sur	mais	tu
être	qui	faire	pouvoir	leur
et	dans	plus	avec	homme
à	en	dire	tout	si
il	du	me	petit	deux
avoir	elle	on	aller	chasser
ne	au	mon	en	heure
je	pour	lui	bien	yeux

Three female and one male French native speakers, all without any articulatory difficulties, were recorded. The four volunteers were asked to read a list of 50 words using two types of articulation: (1) normal vocalized articulation; and (2) silent articulation. It is important to distinguish silent speech from whispered speech. In order to produce silent speech, the speaker is asked to maintain a glottal closure while articulating a word, so that no sound is produced. A crucial aspect of silent speech experiments is to ascertain the importance of the lack of auditory feedback during articulation. Therefore, during all acquisitions, a supervisor verified that subject produced no sound, which was also confirmed by the audio recordings.

[2] http://eduscol.education.fr/cid47916/liste-des-mots-classee-par-frequence-decroissante.html

All recordings began and ended with the lips slightly open and the tongue in a flat rest position. This requirement was used to avoid confusion between bilabials at the beginning and end of the articulation of each word. It was up to each speaker to define and maintain his or her own rest position, without closing the lips.

The speech database acquired contains a total of 12 repetitions of the corpus by each speaker, consisting of 3 vocalized recordings and 3 silent recordings, realized on both the fixed and mobile systems. Within each repetition, the 50 words of the corpus are pronounced individually, in order to allow isolated word recognition. Each word is represented by between 70 and 140 pairs of ultrasound/video images. Thus, altogether, an image database for one speaker on fixed or mobile system contains, on average, 6 recordings × 50 words × ≈210 images/word (105+105=video + ultrasound), for a total of 63,000 images.

4. Recognition algorithm

4.1. Visual feature extraction

First, ultrasound and video images of the tongue and lips are resized to 64x64 pixels. Then, the "EigenTongues" [9] decomposition technique is used to encode each ultrasound frame. In this method, each vocal tract configuration is interpreted as a linear combination of standard configurations, the "EigenTongues", which are obtained after a Principal Component Analysis (PCA) of 5,500 frames, randomly selected from the training set. A similar technique is used to encode frontal images of the lips ("EigenLips"). The number of projections onto the set of EigenTongues/EigenLips used for coding is set to 30 for each visual stream, which accounts for 80 % of the variance observed in the training set [4]. Finally, static visual features are completed by their first and second derivative, so that each tongue/lip image is represented by a vector of 90 coefficients.

4.2. Isolated word recognition

Due to the size of our databases, a Dynamic Time Warping (DTW) technique was adopted for the isolated word recognition step.

4.2.1. Dynamic Time Warping

Dynamic Time Warping is a well-known technique allowing to compare sequences of unequal lengths by dilation or compression along the time axis [10]. In our DTW-based speech recognition system, the DTW is applied, in the visual domain, between feature vector sequence of the test word and feature vector sequence of all the words of the dictionary. In order to combine the two visual modalities, we adopted a "decision fusion" strategy, so that the recognized word $\hat{\mathbf{s}}_{ref}$ is found as:

$$\hat{\mathbf{s}}_{ref} = \arg\min_{k}(DTW(s_{test}, s_{ref}^{k})) \quad (1)$$

where $DTW(s_{test}, s_{ref}^{k})$ is the cumulative distance obtained after a dynamic time warping between the two sequence of visual features of the test word s_{test}, and of the reference word s_{ref}^{k} respectively. Here, a cosine distance is used to measure the similarity between the two vectors. The decision fusion method applies one DTW on tongue features and one DTW on lip features so that:

$$DTW(s_{test}, s_{ref}^{k}) = \lambda \cdot DTW(s_{US,test}, s_{US,ref}^{k}) + (1-\lambda) \cdot DTW(s_{Video,test}, s_{Video,ref}^{k}) \quad (2)$$

In order to find the optimal weighting (the λ parameter) a cross-validation test was performed on one speaker recording set. An optimal 70% tongue (ultrasound), 30% lip (video) weighting was found, which is coherent with [3]. As expected, the tongue carries the most important part of the accessible articulatory information.

4.2.2. Recognition score

Isolated word recognition scores were calculated on the fixed and portable systems in the same way. Four types of comparisons were made for each speaker:

- **Vocalized vs. vocalized**: a word from one of the vocalized repetitions of the corpus is compared with all words in the remaining 2 vocalized repetitions;
- **Silent vs. silent**: a word from one of the silent repetitions of the corpus is compared with all words in the remaining 2 silent repetitions;
- **Two "cross scores"**: a word from one of the vocalized (silent) repetitions of the corpus is compared with all words in the 3 silent (vocalized) repetitions.

In order to increase the statistical relevance of the word recognition performance, a jackknife (leave-one-out) technique [11], in which each list of the recordings was used once as the test set was employed. The performance P for a particular speaker and a given test list is defined by:

$$P = \frac{100 \times N}{L} \quad (3)$$

where N represents the number of correctly recognized words in the test list and L is the number of words in the corpus.

The recognition scores were averaged over the different test lists and speakers. The 95% confidence intervals of these mean values were also calculated by assuming:

$$\Delta_{95\%} = 2\frac{\sigma}{\sqrt{n}} \quad (4)$$

where σ is the standard deviation of P and n is the number of recordings considered for calculating averages.

5. Results

Averaged recognition scores and their 95% confidence intervals appear in tables 2, 3, and 4 for the 4 types of comparisons described in section 4.2.2. Table 2 makes use of the optimal 70% tongue, 30% lip weighting, while recognition results using only tongue, or only lip, data, appear in tables 3 and 4, respectively.

Table 2. *Mean ± 95% confidence interval; results on all speakers using portable system (fixed system results in brackets). Tongue 70%, Lips 30%.*

		Reference	
		Vocalized speech	Silent speech
Test	Vocalized speech	75.3% ± 7.9% (81% ± 3.5%)	42.2% ± 8.8% (30.5% ± 5.4%)
	Silent speech	45.7% ± 7.4% (33.8% ± 7.1%)	64.5% ± 6.5% (65.3% ± 5.8%)

Table 3. *Mean ± 95% confidence interval; results on all speakers using portable system (fixed system results in brackets). Tongue only.*

		Reference	
		Vocalized speech	Silent speech
Test	Vocalized speech	76.2% ± 4.5% (75.8% ± 5.9%)	34% ± 7.9% (25.7% ± 6.1%)
	Silent speech	34.5% ± 7% (27% ± 6.4%)	57.3% ± 6.2% (58.2% ± 6.9%)

Table 4. *Mean ± 95% confidence interval; results on all speakers using portable system (fixed system results in brackets). Lips only.*

		Reference	
		Vocalized speech	Silent speech
Test	Vocalized speech	36% ± 8% (45.8% ± 5.8%)	23.8% ± 5.4% (15.5% ± 5%)
	Silent speech	26.8% ± 6.8% (15% ± 5.3%)	34.7% ± 8.7% (37.8% ± 5%)

Tables 2, 3, and 4 allow us to make a number of interesting conclusions:

- For all choices of tongue/lip weights, and for both silent and vocalized speech, the recognition results are quite similar for the fixed and portable system. This suggests that continued development of more realistic portable systems, for use by healthy and speech handicapped individuals alike, should be quite promising.
- The results using the lips only are very poor, and those obtained using only the tongue are almost as good as those using the optimal weighting. This confirms once again the result obtained, for example in [3], that the tongue information is crucial for an imaging-based SSI.
- The recognition scores obtained when silent speech is compared to a vocalized reference, or vice versa (the cross scores), is significantly worse than that obtained in the vocalized-vocalized case or in the silent-silent case. This concurs with a preliminary result cited in [4] and is of significance for an ultrasound based SSI, since it means that vocalized training data cannot be used directly to train the silent speech recognition system.
- On the other hand, the recognition score obtained when silent speech is compared to a silent reference, though not as good as the vocalized-vocalized case, is significantly better than the cross scores. This implies that it may indeed be possible to devise an SSI using silently acquired training data.

6. Conclusion and perspectives

An isolated word recognition technique applied to ultrasound and video images was used to address two experimental issues that are critical for the design of a realistic ultrasound-based SSI. We have shown that: (1) a portable acquisition system can be used instead of a fixed (and more constrained) one; and (2) that significant differences exist between articulation in normal vocalized speech and in silent speech. Data acquired in vocalized mode will of course be indispensable when a subsequent synthesis step is foreseen. One approach could then be to initially train models on vocalized speech, and then use model adaptation techniques to obtain a model specialized on silent speech.

Clearly, in future work, we will also want to try to understand how silently articulated speech differs from vocalized speech. This will be possible by directly studying articulator movement in the current datasets, and by introducing additional datasets such as vowel-consonant-vowels structures. We will also want to study the effect of "learning" on the results which can be obtained, i.e., can a speaker "learn" how to articulate in such a way as to further improve his recognition score. A real time implementation of the DTW may be one way of helping to attack this problem. Finally, it will be necessary to devise a less massive and more realistic portable system in order to improve user comfort during use.

7. Acknowledgements

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.

8. References

[1] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., and Brumberg, J.S., "Silent speech interfaces", Speech Communication, 52, pp. 270-287, 2009.

[2] Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G., and Stone, M., "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips", Speech Communication, 52, pp. 288-300, 2009.

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

[4] Hueber, T., "Reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal", (PhD, Université Pierre et Marie Curie, Paris, France), 2009.

[5] Stone, M., "A guide to analyzing tongue motion from ultrasound images", Clinical Linguistics and Phonetics, 19(6-7): pp 455–502, 2005.

[6] Whalen, D., Iskarous, K., Tiede, M., Ostry, D., Lehnert-Lehouillier, H., Vatikiotis-Bateson, E., and Hailey, D., "The Haskins optically corrected ultrasound system (HOCUS)", Journal of Speech, Language, and Hearing Research, 48(3): pp 543–553, 2005.

[7] Wrench, A., Scobbie, J., and Linden, M., "Evaluation of a helmet to hold an ultrasound probe", in Ultrafest IV, New York, USA, 2007.

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

[9] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., and Stone, M., "Eigentongue feature extraction for an ultrasound-based silent speech interface", ICASSP, Honolulu, USA, pp. 1245-1248, 2007.

[10] Sakoe, H., Chiba, S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", in IEEE Transactions on Acoustics, Speech, and Signal Processing, pp. 43-49, 1978.

[11] Efron, B., "Nonparametric Estimates of Standard Error - the Jackknife, the Bootstrap and Other Methods", Biometrika 68, 589-599, 1981.