ARTICULATORY STRATEGIES FOR LIP AND TONGUE MOVEMENTS IN SILENT VERSUS VOCALIZED SPEECH

*Lise Crevier-Buchman[a,b], Cédric Gendrot[a], Bruce Denby[c,d], Claire Pillot-Loiseau[a],
Pierre Roussel[c], Antonia Colazo-Simon[a] & Gérard Dreyfus[c]*

[a]Laboratoire de Phonétique et Phonologie, CNRS-UMR 7018, Paris, France;
[b]Voice and Speech Lab, European G. Pompidou Hospital, Paris, France;
[c]SIGMA Laboratory, ESPCI ParisTech, CNRS-UMR 7084, Paris, France;
[d]Université Pierre et Marie Curie, Paris, France
`lise.buchman@numericable.fr`

ABSTRACT

In the context of Silent Speech Communication (SSC) development after total laryngectomy rehabilitation, tongue and lip movements were recorded with a portable ultrasound transducer and a CCD video camera respectively. A list of 60 French minimal-pairs and a list of 50 most frequent French words were pronounced in vocalized and silent mode by one speaker. Amplitude and timing of the articulatory movements were measured and compared in the two modes. This study showed for silent speech, i) a reduced duration of words, ii) a general hypoarticulation for lips, but non significant changes for tongue movements depending on the type of vowel and consonant.

Keywords: silent speech, articulation, labial imaging, ultrasound imaging, tongue movement

1. INTRODUCTION

Silent Speech Communication (SSC) is an emerging technology intended to enable speech communication in the absence of an intelligible acoustic signal [1, 2]. Silent Speech Interface (SSI) devices aim to recognize speech from non-audible signals captured through multisensors placed externally on different parts of the vocal tract. They intend to capture articulatory movements such as lip and tongue movements and resituate appropriate signals for oral communication. Different applications could be proposed such as an alternative voice for total laryngectomized patients. Recent studies [3, 5] have proposed a portable system analyzing in near real time the tongue and lip movements captured by an Ultrasound (US) transducer and a video camera, respectively. It could be interesting to understand on a phonetic level the articulatory differences between normal and SSC speech, to improve recognition performance in the case of SSC, to help and guide rehabilitation to allow patients after total laryngectomy to use efficiently a future SSI after total laryngectomy; to improve recognizer-synthesizer system by better understanding some articulation similarities that are not visible for SSI, and articulatory strategies that are different between vocalized and silent speech; to contribute to a better understanding of some confusion that can occur in SSI and to better train the accuracy of a recognizer-synthesizer system for communication in silent conditions.

The phonetic aspects and specificities of SSC are challenging research fields. Many studies have been conducted to characterize whispered speech [4] but silent speech is a quite new field. In silent speech, the phonological feature [+ voice] cannot be realized at the glottis level as there are no laryngeal activities with any vocal fold vibration. Furthermore, there is no build-up of pressure as there is no airflow passing through the glottis. Therefore, vowel and consonant contrasts depend on articulators such as the jaw, the tongue and the lip movements. [9]. Our study aimed to characterize the difference in duration and amplitude parameters of tongue and lip articulation movements for 44 monosyllabic and 29 disyllabic words in silent and vocalized mode.

2. SPEECH DATA ACQUISITION

For tongue and speech investigation, different ultrasound devices have been described [3]. A portable multisensor system as compared to a fixed one has fewer constraints for the speaker and is closer to the actual situation that would be offered to voice-handicapped patients.

Our subject was a 50-years-old French native female with no known articulation problems.

2.1. Corpus

The non-verbal signals used for SSI such as lips and tongue movements may lead to certain confusion predominantly for labial and lingual look-alike phonemes.

- Therefore we selected a list of 60 minimal pairs with 3 vowel-position in the word (initial, intervocalic and final position such as: [aby] /[oby]; [sil]/[sal]; [kaʁa]/[kaʁo]. Words were represented by either monosyllables with vowel and consonant /V-C/, /C-V/, /CVC/ or disyllables.
- A second list was constructed from the 50 most frequent French words according to the National Education Corpus [8]. In this list we selected 13 monosyllables with isolated vowels such as: [a, e, u, ã, ɛ̃, ɔ̃].

Each list was repeated 3 times using two types of articulation: (1) normal vocalized articulation and (2) silent articulation. Within each repetition, the 60 and 13 words of the corpus were pronounced individually. Each word is represented by between 70 to 120 video frames and 70 to 140 Ultrasound frames. Thus, the image database contains 73x2 words (vocalized + silent) repeated 3 times x 200 images/word (100 video+100 US) for a total of ~ 87,600 images.

2.2. Recording material

An industrial CCD video camera from The Imaging Source, with a 60 fps, fitted with a LED ring light to capture lip movements in a frontal position was used.

The ultrasound (US) system is the lightweight t3000™ developed by Terason, using an 8MC4 microconvex transducer (opening angle: 140°, frequency range: 4-8 MHz) at 60 fps.

The audio signal (16 KHz, 16 bits) was recorded with a head-mounted microphone. It was placed at a constant 5 cm distance from the speaker's lips. This recording was used to confirm the absence of any voicing during silent speech.

A helmet-based system (Fig.1), described in [6] was used for our recordings. The helmet is commercialized by Articulate Instruments Ltd. [10] and was used to stabilize the ultrasound probe to maintain close contact perpendicular to the mouth floor and the hard palate. 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. 1. The system maintains the ultrasound probe and the camera in a stable and fixed position, without restricting head movement.

Figure 1: Helmet fixed on the head with camera mounted in front of the mouth, head-mounted microphone and US fixed under the chin.

Recording started by defining a rest position for the tongue, flat in the mouth, and for the lips slightly opened. Each word started from this rest position and ended the same way. This was to avoid confusion that might have occurred with bilabial consonants in initial or final position.

To prevent speaker fatigue, the acquisition procedure was split into 3 sessions each separated by a 15 min. pause. The helmet was removed and the speaker was able to move around. Each session was represented by the reading of the 2 lists in the 2 modes and lasted about one hour. After each break, before restarting the recordings, a recalibration was done for tongue and lip rest position.

Ultraspeech software: the 3 signals recorded from the US, the video camera and the microphone are controlled by a stand-alone dedicated graphical software interface called Ultraspeech [3], allowing synchronous acquisition of the recorded visual signals along with the acoustic signal. After each acquisition, data are directly available as series of bitmaps for the image streams, and .wav files for the audio.

Ultraspeech provides also an interactive inter-session recalibration mechanism that allows recording of large audiovisual speech databases in multiple acquisition sessions. It was employed to maintain the positioning accuracy of the video sensors (camera and US) across all sessions.

2.3. Visual speech feature extraction

2.3.1. Lips

Movies were recorded for each word and individual pictures were extracted for analysis. A

Matlab program was built-up in order to allow for analysis. On every other frame, points were placed manually in 6 reference places (Fig 2) to measure the external and internal opening, and the stretching-rounding of the lips and euclidean distances were then calculated (in pixels) between these points. Each word was composed of about 70 to 100 frames.

Figure 2: Picture of the lips at rest position taken with the video camera (60 fps), and the 6 points that served to determine on the vertical axis, the inner and extern aperture, and on the horizontal axis, the stretching between the commissural of the lips.


2.3.2. Tongue

Decomposition was used to extract visual speech features from the ultrasound images. In this technique, every second ultrasound image is measured.

The images were measured with the software Edge Track and Matlab. After all reference points from the tongue movements were recorded, data were computed so as to measure the global height and frontness of the tongue for each word.

Figure 3: UltraSound image of the tongue (left) and contours extracted with EdgeTrack software and Matlab program (right).
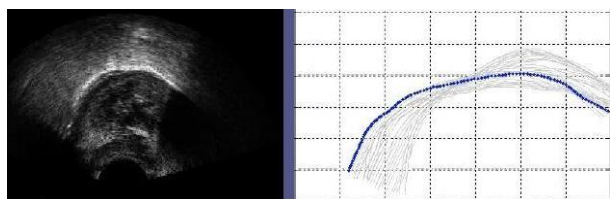
3. EXPERIMENTAL RESULTS

The words of each list were grouped depending on the number of syllables and the structure of the syllable: for the monosyllables, the first group was structured as V, C-V, V-C; the second group of monosyllables was CVC; the third group was represented by disyllables.

Six measures were analyzed: duration of the complete articulation of the word of each list calculated from the number of frames of the words, 3 parameters for the lip and 2 for the tongue, each of them in silent and vocalized mode. They were compared using two-way ANOVAS with mode (silence vs. speech) and words as predictors.

3.1. Monosyllable words (V, CV and VC)

They were statistically longer in the vocalized mode compared to the silent mode (t(1,24)=87, p=0).. The internal and external aperture was wider in vocalized speech (t(1,24)=21.6, p=0).

However, there was en exception for vowels such as [a, e, u, ɑ̃, ɛ̃, ɔ̃] where silent speech had a wider mouth opening, that could represents hyper-articulation. IN addition, vertical tongue movements were significantly wider in silent mode while horizontal movements were significantly wider in vocalized mode. Finally, lip rounding represented by the spread parameter was significantly higher in silent speech (t(1,24)=2.5, p=0.11).

3.2. Monosyllable words (CVC)

They were statistically longer in vocalized mode compared to silent mode (t(1,14)=88, p=0.0). Lips movements were significantly wider in vocalized speech compared to silent speech (t(1,14)=48, p=0.0). No difference was found for tongue movements (t(1,14)=0.8, p=0.37 and t(1,14)=0.008, p=0.92).

3.3. Disyllables words

We found the same results as above with significant difference between vocalized and silent speech for lip movements, with the vocalized words having greater amplitude. No difference was found for tongue movements. Words were statistically longer in vocalized mode compared to silent mode.

4. DISCUSSION AND CONCLUSIONS

The aim of this preliminary study was to identify articulatory strategies in silent speech compared to vocalized speech. The objective was to better understand possibilities and limitations of a SSI and to gain from our results, some recommendations for alaryngeal patients' rehabilitation. After total laryngectomy, there is no more laryngeal vibration and the situation is quite similar to that of silent speech. The absence of voicing features has to be replaced by other features produced in the vocal tract, manly the tongue and the lips in the oral cavity. To represent everyday speech we selected a restricted database composed of frequently used French words and

minimal pairs for look-alike phoneme possible confusion. This database has significant limitation for a structured study at the level of segmental phonetics as we didn't control the vowel and consonant environment and contextual effects. Nevertheless, several conclusions can be addressed.

4.1. Word length

Regardless of word structure in our database, we found a general shortening of word length in silent speech. This is contrary to what has been found for whispered speech [7].

4.2. Lips kinematics

Recent studies on Ultrasound-based SSI interface [2] showed the prevailing importance of tongue compared to lips for an imaging –based SSI. Furthermore, our results showed hypoarticulation movements of the lips in SSC compared to vocalized speech.

The presence of consonants (CVC) in the word seems to stabilize the lip articulation with reduced amplitude of articulatory movement in silent speech; on the contrary, vowels alone showed greater variability in the articulatory pattern when produced silently. One hypothesis could be related to the theory of degrees of freedom [9] depending on the vocal tract shape; in the silent speech situation, the absence of a consonant in a monosyllable could deprive the mental image of the oral cavity during articulation.

Understanding segmental particularities of silent speech should improve the accuracy of articulation rehabilitation for patients after total laryngectomy. Therefore, training patients to improve their articulatory behavior can improve their communication skills as well as the visual signal for SSI.

4.3. Tongue movements

When producing silent speech, in addition to the loss of auditory feedback, there is a loss of aerodynamic information and of intraoral pressure; the consequence is a modification of the adaptation of tongue gestures. For monosyllable (CVC) and disyllables, our results suggest a non-consistent behavior with no significant different movements during silent and vocalized speech. On the contrary, vertical movements are more pronounced in a vocalic or CV context for silent speech. No evident explanation has been proposed yet.

There is still considerable work to be carried out to shed light on silent speech articulatory strategies. A complete set of controlled phonemes would be necessary to outline a typology of this particular mode of communication and a larger population should be tested to circumvent individual variations.

5. ACKNOWLEDGMENTS

6. REFERENCES

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

[2] Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E., Chapman, P.M. 2008. Development of a (silent) speech recognition system for patients following laryngectomy. *Medical Engineering & Physics* 30, 419-425.

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

[4] Higashikawa, M., Green, J., Moore C.A., Minifie, F.D. 2003. Lip kinematics for /p/ and /b/ production during whispered and voiced speech. *Folia Phoniatrica et Logopaedica* 55, 17-27.

[5] Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G., Stone, M. 2009. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication* 52, 288-300.

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

[7] Jovicic, S.T., Sarie, Z. 2006. Acoustic analysis of consonants in whispered speech. *J. Voice* 22, 263-274.

[8] National Education Corpus *http://eduscol.education.fr/cid47916/liste-des-mots-classee-parfrequence-decroissante.html*

[9] Stevens, K. 1999. Articulatory-acoustic-auditory relation-ships. In Hardcastle, W., Laver, J. (eds.), *The Handbook of Phonetic Science*. Oxford: Blackwell, 462-506.

[10] Wrench, A., Scobbie, J., Linden, M. 2007. Evaluation of a helmet to hold an ultrasound probe. *Proc. Ultrafest IV* New York.