DIFFERENCES IN ARTICULATORY STRATEGIES BETWEEN SILENT, WHISPERED AND NORMAL SPEECH ? A PILOT STUDY USING ELECTROMAGNETIC ARTICULOGRAPHY

Thomas Hueber, Pierre Badin, Christophe Savariaux, Coriandre Vilain, Gérard Bailly
GIPSA-lab UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France

Context

The concept of building a device that allows speech communication without the necessity of vocalizing has received considerable attention in the speech research community [1]. Such a device, called a « Silent Speech Interface » (SSI), could be used in situations where silence is required (as a silent cell phone), as well as by laryngectomized patients who can articulate normally but have lost the ability to vocalize speech. The main approaches described in the literature are based on (a) the capture of tongue and lips motion using ultrasound and video imaging, (b) the measurement of the muscle electrical activity using surface electromyography, and (c) the amplification of "non-audible murmurs", using a stethoscopic microphone (NAM). In most of these studies, sensor data are mapped to various speech signal characteristics using statistical models (such as GMM or HMM). These models are trained on multimodal datasets associating articulatory activity with the corresponding audio signal. However, as shown in [2] and [3], the performance of these models trained on "vocalized speech" decreases when they are used to decode "silent speech" (if no model adaptation scheme is applied). This may reveal some differences in terms of articulatory strategies between these two production modes. In this paper, we report preliminary results of a pilot study aiming at characterizing these differences, using electromagnetic articulography (EMA) and acoustic calibration techniques.

Experimental protocol

A native French speaker was asked to repeat twice a list of 160 VCV sequences where V={a ɛ e i y u o ø ɔ œ} and C={p t k f s ʃ b d g v z ʒ m n ʁ l}, in three modes of vocalization: "normal", "whispered" and "silent". For the "silent" condition, the subject was asked to speak as quietly as possible ***while maintaining an intelligible speech production***. To control his very soft production, the subject was given an audio feedback of his own voice through headphones. This feedback signal was captured by a close-talk microphone placed next to the speaker's lips. This signal was also monitored by an "expert listener" who checked the intelligibility of silent speech during data acquisition. In order to measure how "silent" the produced speech was, absolute sound pressure level measurement (SPL) was obtained using a calibrated Brüel&Kjaer microphone placed one meter away from the speaker's face. Articulatory activity was recorded synchronously with the audio signals, using the Carstens 2D EMA system (AG200). Six coils were attached respectively on the tongue tip, tongue blade, tongue dorsum, upper lip, lower lip and jaw. Preliminary tests showed that the presence of the headphone and the close-talk microphone in the EMA system did not alter the accuracy of the measurements.

Results

Mean SPL calculated on all speech segments (manually labeled on the close-talk microphone signal) were 44.0, 47.3 and 62.0 dBSPL for respectively silent, whispered and normal speech while the mean SPL for the sound booth's ambient noise was 43.9 dBSPL. In this study, "silent speech" was thus well defined as an "intelligible" but "non-audible" speech production, *i.e.* indistinguishable from background noise at one meter away from the speaker.

For each of the 320 VCV utterances, an articulatory target for the central consonant C was automatically extracted. This target was defined as the mean articulatory configuration observed when the 3 EMA parameters with the highest peak-to-peak amplitude had reached

their respective extrema. Articulatory targets for the right and left vowels of the VCV utterances were also automatically extracted by determining the most stable configurations before and after the central consonant. Figure 1 shows the dispersion ellipses of the six EMA coils for the 3 vowels {a i u}, for all the consonantal contexts and for all production modes.

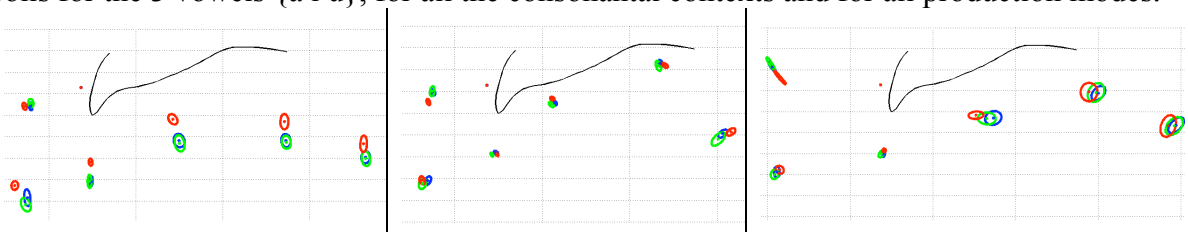

Figure 1: Dispersion ellipses of the 6 EMA coils (lips are on the left side) for 3 vowels {a i u}, in silent (red), whispered (green) and normal speech (blue). The hard palate is represented in black.

For vowels [a] and [u], a first analysis revealed differences between silent and both normal and whispered speech (whereas almost no difference was observed between normal and whispered speech). For vowels [a] (and also to a certain extent for {o ø ɛ ɔ œ}), the position of the jaw and of the tongue was higher in silent speech than in whispered/normal speech. For vowels {u y}, the tongue had more of a frontal position and the lips seemed to be slightly less protruded in silent speech. For vowels {i e}, no difference was observed between the three production modes (except for the back of the tongue for vowel [i]). Concerning the consonants, two tendencies were observed – (1) for the consonants {b v} (but also to a certain extent for {p m n f ʁ l}), the tongue was higher in silent speech than in whispered/normal speech independently from the jaw position (and also further back for {n l}) as shown in Figure 2 - (2) ***silent articulation seems to be more resistant to coarticulation***. In order to quantify this tendency, we calculated the mean area of the dispersion ellipses for all VCV sequences in each production modes; we obtained a mean dispersion of 2 mm^2 in the case of whispered or normal speech, and only 1 mm^2 in the case of silent speech.

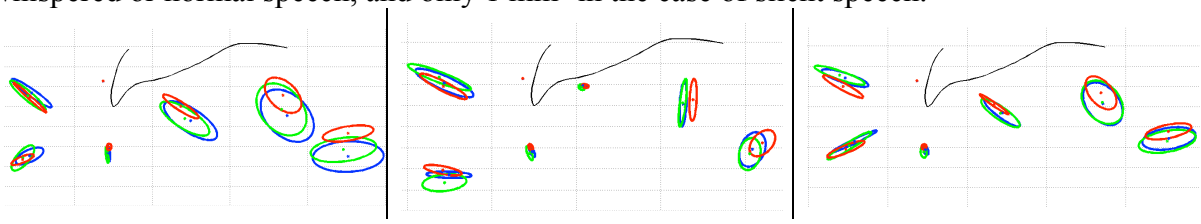

Figure 2: Dispersion ellipses of the 6 EMA coils for 3 consonants {b v n}, in silent (red), whispered (green) and normal speech (blue)

Acknowledgements

The authors would like to acknowledge useful discussions with Maëva Garnier, Nathalie Henrich, Nicolas Ruty, Lucile Rapin and Hélène Loevenbruck.

References

[1] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S. (2009) "Silent speech interfaces", Speech Communication, 52(4), pp. 270-287.

[2] Janke, M., Wand, M., Schultz, T., (2010) “Impact of Lack of Acoustic Feedback in EMG-based Silent Speech Recognition”, Proceedings of Interspeech (Makuari, Japan), pp. 2686-2690.

[3] Florescu, V-M., Crevier-Buchman, L., Denby, B., Hueber, T., Colazo-Simon, A., Pillot-Loiseau, C., Roussel, P. Gendrot, C., Quattrochi, S. (2010), "Silent vs Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface", Proceedings of Interspeech (Makuari, Japan), pp. 450-453.