Effect of a fixed ultrasound probe on jaw movement during speech

Julián Villegas[1], Ian Wilson[1], Yuki Iguro[1], and Donna Erickson[2]
[1]University of Aizu, Japan, [2]Kanazawa Medical University, Japan

Abstract
The use of an ultrasound probe for observing tongue movements potentially modifies speech articulation in comparison with speech uttered without holding the probe under the jaw. To determine the extent of such modification, we analyzed jaw displacements of three Spanish speakers speaking with and without a mid-sagittal ultrasound probe. We found a small and not significant effect of the presence of the probe on jaw displacement. Counterintuitively, when speakers held the probe against their jaw larger displacements were found. This could be explained by a slight overcompensation on their speech production.

Method
We recorded three native speakers of Spanish uttering seven repetitions of 26 sentences (7 in English, 3 in Japanese, and 16 in Spanish) with and without the ultrasound probe fitted under their chin for a grand total of 1,092 sentences. For the statistical analysis, we used all sentences we recorded excepting those that had some capture (or trace extraction) problems. In total, 912 sentences were used in the analysis (252 in English, 107 in Japanese, 553 in Spanish).

Speakers
The three female speakers (s1, s2, and s3) were Salvadoran of 23, 28, and 34 years of age, with varying degree of second and third languages exposure: while the eldest reported ten years of English studies and three of Japanese (she had lived the last six years in Japan), s1 and s2 reported five and ten years of English training. None of these two had Japanese training. The youngest speaker had also lived in the USA for one year, immediately preceding the data collection whereas s2 had lived mainly in El Salvador. With the exception of s1, the speakers reported to still have a neutral Salvadoran Spanish accent, as acknowledged by their Salvadoran acquaintances and relatives.

Materials
The sentences were selected so the same vowel was prominently used in all the constituent words. These sentences are summarized in Appendix 1. A tripod-mounted Panasonic HDC-TM750 digital video camera was used to collect video of the front of the face. Light from two 300W halogen bulbs (LPL-L27432) was reflected onto the face to improve automatic marker tracking. Audio was recorded with a DPA 4080 miniature cardioid microphone connected to a Korg MR-1000 digital recorder, and tongue movements were recorded with an ultrasound Probe - Toshiba (PVQ-381A) connected to an ultrasound machine - Toshiba Famio 8 (SSA-530A).

Procedure
Speakers were recorded in two sessions: first without and second with an ultrasound probe under their chin. Each session was comprised of three blocks corresponding to the three languages recorded in this order: Spanish, English, and Japanese. Each block of utterances was randomly sorted and presented from a laptop computer located at about two meters in front of the speaker in Calibri black font (44 points) over white background. We prevented head tilting by changing the height of the display for each participant. Errors (mainly coughs, reading errors, and ultrasound probe misalignments) were marked visually and aurally in the video and audio recordings, prior to having the speaker repeat the dubious token.

Speakers were able to take short breaks between blocks and sessions. The two sessions were recorded in about one hour. Permission for performing these recordings was obtained following the University of Aizu ethics procedure.

After instructing the speakers about the experiment and querying them about their language background, they were asked to sit straight in a well-lit room, in front of a white background. The experimenters (two in each session) assisted them with putting on a lapel microphone. Subjects also don a glasses frame (without lenses) with a blue circle of about 8 mm in diameter, located at the center of the frame, above the participant's nose; a second marker was placed by the experimenters on the chin of the speaker and perpendicular to the frame line, as shown in Figure 1. Speakers were recorded in video at 29.97 frames per second (i.e., samples were taken every 33.367 ms) and at 44.1 kHz/16 bits in audio.

Figure 1. A speaker speaking without (left) and with the ultrasound probe (right). One marker (blue dot) was located on the lensless glass frame while the second marker was placed on the speaker's chin.

Post-processing

End points of each utterance were located from the audio of the video recordings in Praat [1] by visual inspection. These end-points were used to extract the videos using ffmpeg routines (https://ffmpeg.org). From the extracted videos, the blue dots were traced using the marker tracker program described in [2]. These trajectories were used to compute the Euclidian distance between the markers. Conversion from pixels to mm was approximated by measuring the physical frame (133 mm) and its corresponding videotaped counterpart (398 pixels).

Results

Each token was time normalized (*normT* – dividing each sample time by the length of the sentence) before fitting a smoothing cubic spline ANOVA (SSANOVA) model as implemented by Gu [3]. Note that this method has been successfully used in similar analyses such as F0 contours and larynx height for Mandarin tones [4] and the lingual and labial articulation of whistled fricatives [5].

In our model, jaw displacement (*distance*) is explained by the factors *Probe* (yes or no), *Sentence* (as in Appendix 1), *normT*, and the interaction between the two last factors. As a sole random factor we used *Speaker* (s01, s02, s03). We also used a Generalized Cross-Validation method for smoothing (as implemented in the SSANOVA library) with the default alpha value (i.e., $\alpha = 1.4$). The resulting model has an $R^2 = .484$, with no apparent redundancy on the fixed factors, and a relatively large variability explained by the random factor. This last

finding was expected since subjects had a large variability in jaw opening per sentence repetition (especially when speaking in unknown or poor proficiency languages).

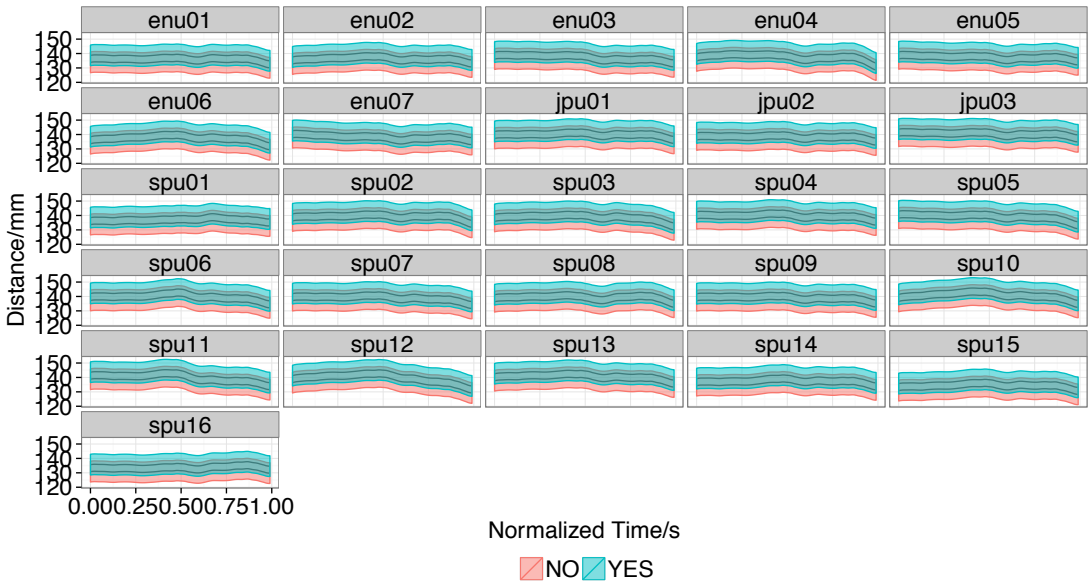

Figure 2. Time contour of the jaw opening for each of the studied sentences. Contours are plotted with their corresponding 95% Bayesian confidence intervals (CIs). Overlapping CIs suggest non-significant differences.

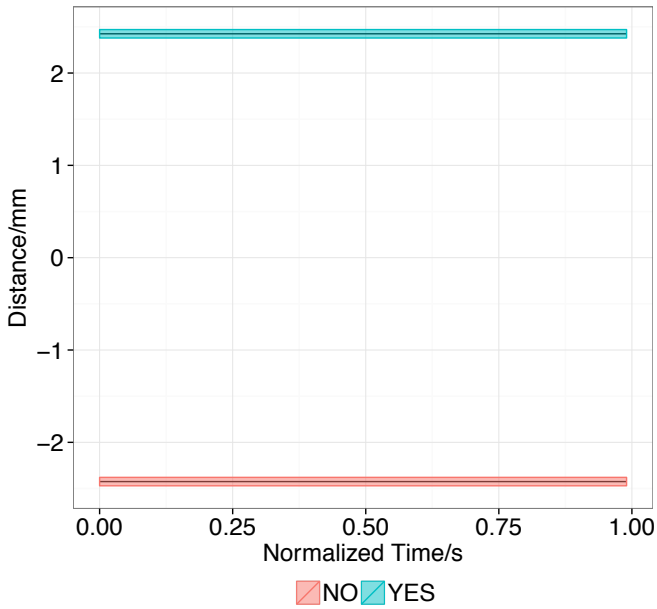

Figure 3. Distance difference predicted by the SSANOVA model for subjects holding the probe under their chin (YES) and when no probe was used (NO).

Findings

The resulting splines per sentence are presented in Figure 2. Interestingly, on average, subjects opened the jaw more when holding the probe under their chin than when they had no probe. When all sentences are considered, this difference was of about 5 mm as shown in Figure 3. The distance between markers varies with subject (i.e., larger subjects exhibit larger distances); in our case, speaker s2 had the smallest distances among the speakers; this is

reflected on the negative offset associated in the model (-5.976 compared to 0.0567 and 5.919 mm for speakers s1 and s3).

Conclusions

We did not find evidence supporting that the presence of an ultrasound probe located under the chin on the mid-sagittal plane hinders the jaw movement of the speakers. The small effect that we found was not significant and in opposition to the expected direction: i.e., it suggests that when the probe was present, subjects were opening the jaw more, probably as an overcompensation reaction.

Acknowledgements

This work was partially supported by the Japan Society for the Promotion of Science (JSPS), Grants-in-Aid for Scientific Research (C) #25370444.

References

[1] P. Boersma and D. Weenink. Praat. Available [Nov. 2015] from www.praat.org
[2] Barbosa, A. V., and Vatikiotis-Bateson, E. Video tracking of 2D face motion during speech. In *Signal Processing and Information Tech., IEEE International Symposium on* (pp. 791−796). (2006)
[3] Gu, C. (2014). Smoothing Spline ANOVA Models: R Package GSS. *J. of Statistical Software*, 58(5):1–25.
[4] Moisik, S., Lin, H., and Esling, J. (2013). *Larynx Height and Constriction in Mandarin Tones*, volume Eastward Flows the Great River: Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday, pages 187–205. City University of HK Press.
[5] Lee-Kim, S.-I., Kawahara, S., and Lee, S. J. (2014). The 'whistled' fricative in Xitsonga: its articulation and acoustics. *Phonetica*, 71(1):50–81.

Appendix 1

Sentences used on the experiment

SentenceID	Sentence
spu01	Ojo con los Orozco: Nosotros no somos como los Orozco, yo los conozco, son ocho los monos: Pocho, Toto, Cholo, Tom, Moncho, Rodolfo, Otto, Pololo. Yo pongo los votos sólo por Rodolfo, los otros son locos, yo los conozco, no los soporto.
spu02	Mamá va a trabajar a Casablanca mañana.
spu03	¿Mamá va a trabajar a Casablanca mañana?
spu04	Ana va a trabajar a Casablanca mañana.
spu05	¿Ana va a trabajar a Casablanca mañana?
spu06	Mamá va a trabajar a la Casa Blanca mañana.
spu07	Ana va a trabajar a la Casa Blanca mañana.
spu08	Mamá valsará "Casablanca" mañana.
spu09	Ana valsará "Casablanca" mañana.
spu10	Mamá valsará "a la Casa Blanca" mañana.
spu11	Ana valsará "a la Casa Blanca" mañana.
spu12	Papá pasará la cámara mañana
spu13	Me elevé en el edén
spu14	Difícil vivir sin ti
spu15	Toto no podó los chopos
spu16	Tu tutú susurra
jpu01	Aka pajama da.
jpu02	Aka gasa da.
jpu03	Dakara Mana wa atamaga sarasara da.
enu01	That fat cat sat with Pat on the mat.
enu02	Pat the fat cat on the mat.
enu03	Pat, the cat, and the gnat sat on the mat.
enu04	That bad, bad cat sat back fast.
enu05	I saw lights in the sky tonight.
enu06	I saw five bright highlights in the sky tonight
enu07	I gave dates to Kay today