EXTRACTING KINEMATIC PROPERTIES FROM LARGE-SCALE ULTRASOUND CORPORA

Martine Toda

EXTRACTING KINEMATIC PROPERTIES FROM LARGE-SCALE ULTRASOUND CORPORA

Martine Toda

Laboratoire de Phonétique et Phonologie, UMR 7018 - CNRS/Université Paris III, France
martinetoda@gmail.com

ABSTRACT

This study describes a highly automated approach to quantifying articulatory movements in ultrasound echography, and offers an example of its application in the examination of the timing of consonant-vowel coordination in typologically contrasted languages (French and Japanese).

Despite recent developments in data acquisition and storage facilities, ultrasound requires a time-consuming post-process consisting of e.g., semi-manual tongue contour tracking, that prevents it from being fully involved in large-scale phonetic corpora.

To take over this limitation, our approach looks at frame-to-frame pixel-wise variation of brightness. The articulatory targets can then be expressed as local minima of the rate of change.

Preliminary results from a bilingual speaker suggest a difference in the time coordination of consonantal and vocalic targets in French and Japanese before and after the intervocalic /k/.

Keywords: ultrasound, large-scale phonetic study, consonant-vowel coordination, image processing

1. INTRODUCTION

1.1. The ultrasound technique

Ultrasound echography is a very attractive technique in acquiring articulatory data of the tongue: there is no health risk; it is easy to use; recording sessions are not technically limited in time, etc. It therefore has high potential to be combined with other articulatory or physiological modalities, e.g. [6], and for large-scale corpus studies.

Ultrasound has important limitations, however, that need to be addressed: (1) low time resolution; (2) only tongue contour (especially tongue dorsum) is visible, not the palate; (3) images need some processing in order to be interpreted in articulatory terms.

Because of its low time resolution (around 30 frames per second, which is much lower compared to electropalatography or EMA – 100 to 200 Hz), ultrasound is often used for the observation of tongue shape at phoneme targets, rather than for the study of articulatory timing. In order to have a clear idea of the place of articulation and degree of constriction, the palate contour is important.

Many attempts have therefore been made in order to register the palate contour in the tongue contour images [5, 7]. However, those systems render recording conditions less comfortable for the subjects or the setup is technically more complex, and the recording of parallel modalities (oral airflow, intraoral pressure, electro-glottography, fiberscopy…) is difficult.

Considering that ultrasound is the only known method for observing tongue motion in a large-scale perspective, the aim of this paper is to propose a highly automated method to extract the kinematic properties of the tongue during speech. Although the time resolution of this technique is low, we assume that the consistent kinematic features should emerge from the data if their volume is sufficiently large. Moreover, as far as kinematic properties are observed, the palate reference is less crucial. Therefore, this approach allows ultrasound to become suitable for the investigation of large corpora.

1.2. Language typology and coarticulation patterns

Two typologically contrasted languages, French and Japanese, were investigated in order to illustrate the proposed method. These languages differ by their consonantal (C) system, vocalic (V) inventory, and their syllable structure. Regarding stops, French and Japanese distinguish labial, coronal and velar places.

In Japanese, the plain/palatal contrast is found in syllable onsets (/p-p^{j}/, /t-t^{j}/, /k-k^{j}/, /b, b^{j}/,…), while C clusters are limited to homorganic nasal-stop sequences or geminates.

In French, complex C clusters are allowed, up to 3 in word initial position (e.g. /stri/ ‘*strie* (stripe)’) and final position (e.g. /arbr/ ‘*arbre* (tree)’or /astr/ ‘*astre* (star)’).

To sum up, in Japanese, C tend to be systematically surrounded by V or a nasal, thus a cue for the C place of articulation is consistently found in the formant transitions adjacent to C. Furthermore, the plain/palatal contrast in Japanese relies on transitional information as well, so that V trigger C cues in a consistent way. In contrast, C place cues are

not necessarily given by formant transitions in French.

Let us suppose that Japanese speakers intend to make the transitional information more salient at the vowel edges. If this were true, it can be hypothesized that the articulatory transition from consonant to vowel or from vowel to consonant will be made slower or the movement peak will be shifted towards the vowel target, just as the formant transition on the spectrogram is steeper in French /ta/ than in /tja/, while the locus and targets are the same in the both. To address this issue, the relative timing of consonant and vowel targets and transitions will be examined.

2. METHOD

2.1. Data collection

One bilingual speaker of French and Japanese participated in this pilot experiment. If there is any reliable difference observed between Japanese and French conditions, it would be related to language, not to the subject's anatomy.

The subject produced eight repetitions of /kVkV/ utterances (phonotactically legal in both languages) where V=/a, i, o/; embedded in the frame sentences "*Je dis ... clairement*" for French and "*hakkirito ... toiu*" ("I say … clearly") for Japanese conditions. All the stimuli were real words or morpheme, as e.g. "利き /kiki/" in "利き手 (one's dominant hand)". In Japanese, unaccented lexical items were chosen, so that the F_0 contour was close to the realization of French items. Also, /o/ was used as a reference back vowel instead of /u/ because there is more phonetic similarity between the two languages for /o/ than for /u/. The velar consonant was chosen because its place of articulation is not masked by the maxilla shadow as in coronal place. The whole corpus was repeated four times by alternating Japanese and French, leading to a total of 32 repetitions for each token in each language.

The ultrasound video was recorded with a Mindray 6600 scanner and a microconvex probe at 29.97 frames per second (NTSC standard). The analogue video output was acquired on a computer via a video card, with the Articulatory Assistant Advanced (2.12) software (Articulate Instruments). The ultrasound probe was held by the subject, who aimed to minimize its rotation during the recordings, but allowed for the jaw to move naturally. Therefore, the rapid (say, > 5 Hz) and systematic displacement of bright areas in the images could reasonably be interpreted as related to the intrinsic movements of the tongue, and not to the possible movements of the probe, which is typically slower and random, and assumed to fade out over the long run.

The speech signal was recorded at 44.1 kHz with an AKG C520 headset microphone, via an external Roland Edirol UA-25ex sound card on the same computer.

2.2. Automatic detection of acoustic events

In order to get reference points for acoustic-articulatory comparisons, the acoustic signal was annotated. The /kVkV/ tokens were labeled manually in Praat for this preliminary study. This procedure can be replaced by an automatic alignment (e.g. [2]) in a larger corpus.

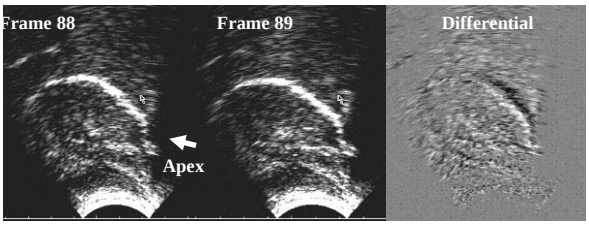
Then, a template matching algorithm (a custom Matlab program) was run for each token. After the signal was down sampled by 10 in order to speed up the process, the frontiers of acoustic segments were located by looking at the abrupt change of zero-crossing rate (ZCR) and energy envelope, parameters that are known to be appropriate in classifying silence, noise and voiced intervals [1].

The algorithm looked for a burst (increase of energy: peak of its derivative), a vowel (decrease of ZCR followed by an increase of energy), a silence (decrease of energy), and again, a burst, a vowel, and silence. About 10% of the alignment (of randomly chosen samples) was checked manually. Only the onset of very weak bursts was not properly detected, but otherwise the landmarks were correct.

2.3. Quantifying articulatory movements

The ultrasound video was converted into as many images as video frames. Tongue movements appear on the ultrasound images as displacement of bright-colored areas (Figure 1, left and center).

Figure 1: Ultrasound images (left and center) during the transition from /k/ to /o/ in the second syllable of /koko/ in Japanese condition. The image differential is shown at right. Tongue contour as well as details in deeper layers moved backward from black to white areas. Medium gray denotes small or no change between the two images.



In many studies, tongue contour is detected (e.g. [4]) prior to further analysis. Tongue contour tracking should be used with caution, however, as the visible contour does not necessarily correspond to a constant physical tongue segment (apex is hidden and reappears), tongue contour sometimes is

discontinuous, or there could be two possible contours (during palatal contact). Therefore, manual correction is often necessary, making this process time-consuming and possibly inaccurate.

An alternative approach was proposed in [3], that extracts eigen components (eigentongues) of the image in a fully automated manner. However, this method can be sensitive to the displacement of the tongue within the imaged area, and thus the probe needs to be fixed with respect to the subject's head.

In order to allow for small discrepancies in tongue position with respect to the ultrasound probe from one repetition to another, while keeping the processing automatic and reliable, we propose a method that captures the differential of the images (illustrated in Figure 1, right). Given that large tongue displacements will result in more extreme values (black and white areas) and less medium values (gray) in comparison with static intervals, the rate of change (RoC) between any two images can be formulated as follows:

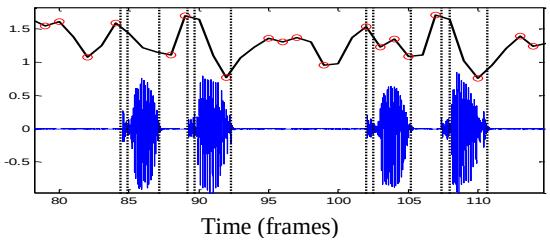
$$(1) \qquad RoC_i = \sum_{j=1 \text{ to } 480 \,;\, k=1 \text{ to } 640} (b_{i,j,k} - b_{i-1,j,k})^2$$

where *b* denotes brightness; i, frame number; and j and k are vertical and horizontal coordinates of pixels.

This method presents the advantage of being very quick to process compared to eigentongues. It also takes advantage of the movements of small structures in the deeper layers of the tongue, that are redundant with that of tongue contour.

As shown in Figure 2, the RoC exhibits local maxima that correspond to larger tongue displacements during transitions, and local minima corresponding to articulatory targets. In the first /koko/ utterance of this figure, the initial /k/ target is reached at frame 82, the intervocalic /k/ has a target at frame 88, and the final /o/ has a target at frame 92. In the second utterance, the first /o/ does have its own target at frame 103.

Figure 2: Speech signal (bottom) with automatically detected segment boundaries, and rate of change (RoC; above, rescaled by $3*10^{-6}$) in function of ultrasound frame intervals (Japanese, V = /o/). Local maxima and minima are highlighted by small circles.

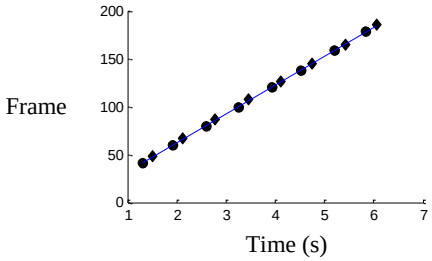


3. RESULTS

3.1. Estimation of real frame rate based on acoustic and articulatory events

We would like to bring to the reader's attention to the fact that the alignment of speech and video signals is not necessarily accurate at this stage. From visual inspection of the signals, the synchronization was good enough so that utterance to utterance correspondence between acoustic and articulatory signals was correct.

Figure 3: Scatter plot of V1 (circles) and V2 (diamonds) onsets detected in the speech signal against the movement peak (RoC maxima) derived from ultrasound, during the k1 to V1 and k2 to V2 transitions, respectively (French, V=/a/). The line represents the best fitting first-order polynomial.



In order to estimate the accurate frame rate of the video with respect to sound, we used the onset of V1 and V2 of each token as acoustic landmarks and the corresponding vowel targets (RoC minima) as articulatory landmarks. The slope (i.e. frame rate) was estimated by fitting a first order polynomial (Figure 3). From a total of 14 sentences where the /kV/ transition peaks were detected for at least 4 utterances, giving a grand total of 111 valid utterances, the average frame rate was 29.99 and its standard deviation 0.12 frames/s.

3.2. Consonant-vowel coordination in French and Japanese

The time lag (duration) between the adjacent phoneme targets (RoC minima) was calculated (Figure 4).

Because of a larger number of utterances with /i/ where no V1 target peak was found, the number of valid utterances for this context is smaller than in the other vowel contexts.

The results show a tendency for /kVkV/ utterances in the Japanese condition to have a longer VC interval than in French, whereas the CV intervals are shorter.

Figure 4: Average duration of k1 to V1, V1 to k2 and k2 to V2 targets for French (diamonds) and Japanese for V = /a, i, o/. Small bars indicate standard error.

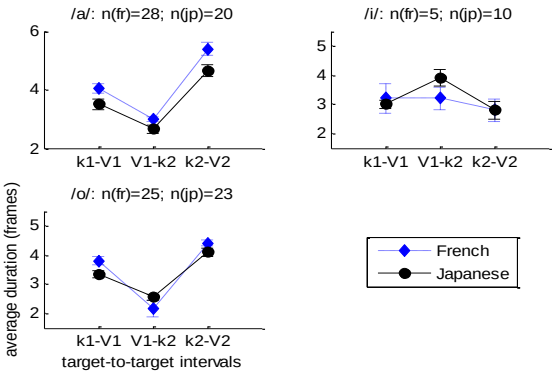


Figure 5: Scatter plot of the normalized average position in time (in %) of transitional peaks with respect to the preceding and following phoneme targets. X axis: French; y axis: Japanese

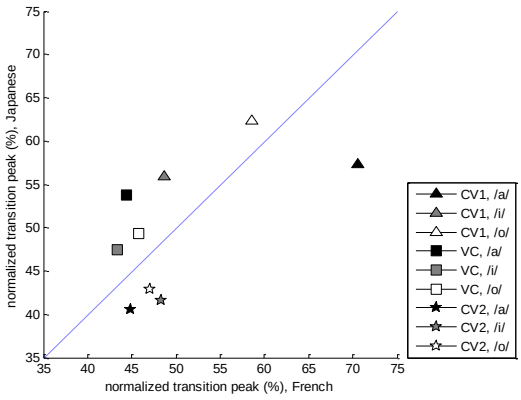


Figure 5 shows the normalized position of transitional peaks with respect to the preceding and following phoneme targets. 0% corresponds to the position of the preceding target, and 100% to the following target. The V1 to k2 transitions (squares) and k2 to V2 transitions (stars) exhibit a language-related difference. Some conditions reach significance: /k1a1/ ($p<0.0001$); /a1k2/ ($p<0.05$); but also /k2a2/ ($p<0.2$) and /k2o2/ ($p<0.2$), according to Student's t test. While the position of transitional peak is constantly located around 47 % of the interval in French, independently of CV or VC transitions, in Japanese, it tends to be shifted towards the consonant in both VC (around 51%) and CV transition (around 40%).

4. DISCUSSION AND CONCLUSION

This preliminary study used a highly automated method to detect articulatory events from standard frame rate ultrasound video of the tongue. It appeared that a language-specific coordination pattern of consonant and vowel targets and transitions exists in the production of /kVkV/ sequences by a French/Japanese bilingual subject. The results are however surprising, as the shift of the CV transition peak towards the consonant is counterintuitive with respect to our hypothesis. Further analysis is necessary in order to clarify the articulatory correlates of this kinematic pattern.

Still, the observed language-related tendencies indicate that the proposed method is able to capture some relevant speech parameters. It opens new perspectives in the corpus study with ultrasound, particularly in examining the coarticulation patterns of languages, but also of the individual subjects. For example, in speech inversion, the articulatory models could be adapted to the speaker on the basis of his/her kinematic properties monitored by means of, e.g., a portable scanner, so that the performance in acoustic-to-articulatory inversion could be improved by applying subject-specific constraints.

5. ACKNOWLEDGEMENTS

This work was supported, in part, by the French National Research Agency (ARTIS project).

6. REFERENCES

[1] Atal, B.S., Rabiner, L.R. 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE TASSP* 24(3), 201-212.

[2] EasyAlign. *http://latlcui.unige.ch/phonetique*

[3] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M. 2007. Eigentongue feature extraction for ultrasound-based silent speech interface. *ICASSP*, 1245-1248.

[4] Li, M., Kambhamettu, C., Stone, M. 2005. Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics* 19(6), 545-554.

[5] Stone, M., Davis, E.P. 1995. A head and transducer support system for making ultrasound images of tongue/jaw movement. *J. Acoust. Soc. Am.* 98(6), 3107-3112.

[6] Vaissière, J., Honda, K., Amelot, A., Maeda, S., Crevier-Buchman, L. 2010. Multisensor platform for speech physiology research in a phonetics laboratory, *Journal of the Phonetic Society of Japan* 14(2), 65-77.

[7] Whalen, D.H., Iskarous, K., Tiede, M.K., Ostry, D.J., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E., Hailey, D.S. 2005. The Haskins optically corrected ultrasound system (Hocus). *JSLHR* 48, 543-553.