Interspeech 2018
2-6 September 2018, Hyderabad, India

Analyzing vocal tract movements during speech accommodation

Sankar Mukherjee[1], Thierry Legou[2], Leonardo Lancia[2], Pauline Hilt[1], Alice Tomassini[1], Luciano Fadiga[1,3], Alessandro D'Ausilio[1,3], Leonardo Badino[1], Noel Nguyen[2]

[1]Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy
[2]Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[3]Section of Human Physiology, University of Ferrara, Italy
sankar1535@gmail.com, noel.nguyen-trong@univ-amu.fr

Abstract

When two people engage in verbal interaction, they tend to accommodate on a variety of linguistic levels. Although recent attention has focused on to the acoustic characteristics of convergence in speech, the underlying articulatory mechanisms remain to be explored. Using 3D electromagnetic articulography (EMA), we simultaneously recorded articulatory movements in two speakers engaged in an interactive verbal game, the domino task. In this task, the two speakers take turn in chaining bi-syllabic words according to a rhyming rule. By using a robust speaker identification strategy, we identified for which specific words speakers converged or diverged. Then, we explored the different vocal tract features characterizing speech accommodation. Our results suggest that tongue movements tend to slow down during convergence whereas maximal jaw opening during convergence and divergence differs depending on syllable position.

Index Terms: Speech Convergence, Dual EMA, human-human interaction.

1. Introduction

During verbal interaction two individuals become part of a complex system whose information flow is mediated by visible behavior, prior knowledge, motivations, inferences from the partner's mental states, and history of prior interactions [1]. While interacting, they adjust their speech to accommodate to each other [2]. Within accommodation, we can identify Convergence (when speakers' speech characteristics become progressively more similar) and Divergence (when speakers move away from the speech characteristics of each other).

Research in this area has dealt with the quantification of speech convergence via objective acoustic measures [3] or subjective evaluations [5] [6], showing a great deal of inconsistency [7]. Part of the complexity is probably due to its dependency on contextual, social and linguistic factors. Previously, our group devised a robust method to extract phonetic convergence in a game-like speech turn-taking task, by using a speaker verification technique [8] [9].

In parallel, the investigation of convergence at the articulatory level is still sparse. For this purpose, we evaluated articulatory dynamics underlying speech accommodation. We asked pairs of participants to engage in an interactive speech task [9] while dual-EMA was recorded. Our first aim was to apply the same technique used in [9] to verify its robustness on a different data-set. More importantly though, we intended to investigate what happens in the articulatory features space, during Convergence and Divergence.

2. Materials and method

2.1. Domino task

We asked native French speakers to perform a Verbal Domino Task (VDT) [8][9] with French words (Fig. 1B). VDT consists in two speakers taking turn in chaining disyllabic words according to a rhyming rule. After listening a word pronounced by a participant, the other participant must choose between two alternative words, such that the first syllable matches the last syllable produced by his/her partner in the game. To build the word chain, we first selected disyllabic words from the Lexique-3 (http://www.lexique.org/) French lexical database. This database was manually checked to exclude crude or offensive words. The chain was built by using a custom made iterative algorithm, which started from the highest frequency word and then looked for the next highest frequency item, fulfilling the rhyming criteria and no repetitions. In this manner, we generated sequences of 300 unique disyllabic words for the VDT.

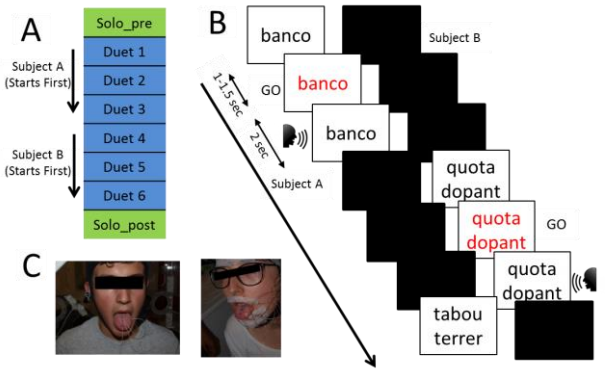

Figure 1. *(A) Experimental timeline including the Solo_pre, Solo_post and duet sessions. (B) Sequence of events during the VDT. (C) EMA sensor positioning in one participant.*

10.21437/Interspeech.2018-2084

2.2. Participants

Participants were ten healthy right handed native French subjects (6 males and 4 females, age range 21-26), who did not know each other, composing 5 same-gender dyads. Everyone had self-reported normal hearing. All participants gave informed consent to participate to the experiment. Procedures were approved by the Ethics Committee of the Ferrara University in accordance with the ethical guidelines of the Declaration of Helsinki.

2.3. Procedure

The experiment was divided into three main sections (Fig. 1A). Solo recordings were performed before and after the Duet sessions (Solo_Pre, Solo_Post). The Solo required subjects to read 60 words to establish a subject-wise baseline. These words were phonetically balanced and selected from the 300 words chosen for the VDT. During the Solo the other participant was listening to classical music through headphones.

During the Solo, one word at time was presented on a black screen and after a variable delay of (1 - 1.5 s) a GO signal instructed the subject to read it aloud. This random delay was introduced to avoid anticipation and entrainment to the rhythm of presentation. For the same reason, trials presentation was intermingled with random delays (2-2.5 s). Since each subject completed 60 words in the Solo, we collected a dataset of 1200 words. The Solo sessions lasted about 4 minutes.

In Duet, the task started with one word presented on the screen of one subject (Subject A), while the other participant's screen was blank (Subject B). Subject A waited for the GO signal (delay of 1-1.5 s) and had 2 seconds to respond. At the end of the trial, Subjects A's screen went blank and two words appeared on Subject B's screen. Now, Subject B had to choose which word to read aloud as only one was complying with the rhyming rule. This chain of events continued until the end of the list.

The 300 words of VDT were divided into 3 lists of 100 and repeated twice so that the Duet part was composed by six separate sessions. In each session, the two speakers read 50 words each, summing up to 300 words per speaker and thus resulting in a total of 3000 words. The duet sessions lasted about 30 minutes. The VDT was implemented in a Psychotoolbox 3 script running in the Matlab environment.

Speech was recorded by two high-quality microphones (AKG C1000S) and the speech data were digitized and acquired by an acquisition CPU (16 bit, stereo, 22050Hz sampling frequency). Both signal went through an external dedicated amplifier (MMX-11USB 2ch audio mixer) and acquired with a A/D acquisition board (MC measurement computing USB-1608GX-2AO).

Articulatory data was recorded with two EMA systems. The first one was an NDI (Northen Digital Instruments, Canada; sampling frequency, 400 Hz) and the second one was an AG501 (Carstens Medizinelektronik GmbH; sampling frequency, 256 Hz). Seven 5-degrees-of-freedom (5-DOF, x,y,z, pitch and roll) sensor coils were glued on the Upper Lip (UL), Lower Lip (LL), Upper Incisor (UI), Lower Incisor (LI), tongue tip (TT), tongue middle (TD) and tongue back (TB). For head movement correction, a 6-DOF sensor coil was fixed on the bridge of a pair of glasses worn by the participants (Fig. 1C).

3. Pre-Processing

3.1. Acoustic Pre-Processing

Incorrect trials (e.g., wrong pronunciation, wrong choice of words, about 3.1%) were excluded from the analysis. Periods of silence were discarded using an energy-based Speech Activity Detector. We then computed MFCCs (Mel Frequency Cepstral Coefficients) by segmenting the data into 25ms frames (10ms overlap) with a Hamming window. The short-time magnitude spectrum, obtained by applying FFT, was passed to a bank of 30 Mel-spaced triangular bandpass filters, spanning from 0 Hz to 3,800 Hz. The output of the 30 filters were transformed into 12 static, 12 velocity and 12 acceleration MFCCs with the 0'th coefficients resulting in 39 MFCC dimensions in total.

3.2. Articulatory Pre-Processing

Articulatory data from both EMA systems, was down-sampled at 100Hz. We removed from the dataset all words for which one or more sensors were detached (Convergence: 36.14±23.36%; NoChange: 32.27±16.44%; Divergence: 42±20.5%). Vocal articulator trajectories (x, y, z positons of the sensor coils) were filtered using an adaptive median filter (10-50ms window) and further smoothed using a 20Hz cutoff elliptic low-pass filter. Coils rotation was ignored. From the x-y midsagittal coil positions we extracted six vocal tract features: lip aperture (LA) (equation 1), lip protrusion (PRO) (equation 2), jaw opening (JO) (equation 3), tongue tip constriction degree (TTCD), tongue blade constriction degree (TBCD) and tongue dorsum constriction degree (TDCD).

$$LA = |ULy - LLy| \quad (1)$$

$$PRO = |ULx - LLx| \quad (2)$$

$$JO = |UIy - LIy| \quad (3)$$

TTCD, TBCD and TDCD are the Euclidean distance of TT, TB and TD to the curve of the palate on the midsagittal plane. To assess how fast these vocal tract features are changing, the velocity of these features was also computed. Since words in the VDT are disyllabic, we expected two local maxima for each word in the jaw opening trajectory, which roughly correspond to the open configuration of the vocal tract for the two vowels. For this reason, we computed the maximum jaw opening of the two syllables separately (JO_Syl_1; JO_Syl_2), and their average (JO_Syl_1&2).

3.3. Convergence and Divergence calculation

To extract an un-biased measure of Convergence and Divergence, we used a data driven, text independent, automatic speaker identification technique [9], based on Gaussian Markov Modelling (GMM) Universal Background Model (UBM). The Gaussian components model the underlying broad phonetic features (i.e. MFCCs) that characterize a speaker's voice. We used the Sidekit-Python library for GMM-UBM modelling. UBM was trained with the pooled Solo_pre speech data of all the participants. Then, individual speaker-dependent models were obtained via maximum a posteriori (MAP) adaptation of the UBMs to the Solo_pre speech data of each speaker separately.

A cross-validation technique was used to choose the optimum number of GMM components. Solo_post speech was used as a validation set, and each speaker-dependent model performance was verified against the UBM model. Finally, a

256-component GMM was chosen as it had the lowest Equal error rate (EER) and showed a good modelling performance of the confusion matrix for the cross validation set (EER=4%; Fig. 2).

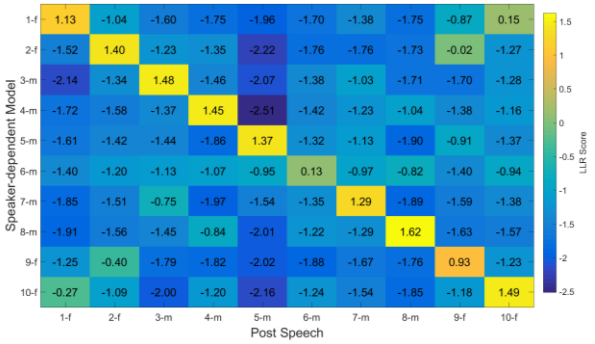

Figure 2. *Speaker verification confusion matrix of all the speaker-dependent models against background UBM in the Solo_Post. Here the diagonal positive score line indicates a good MAP adaptation. The numbers are the subtraction between speaker-dependent model and UBM model and is represented by Log-Likelihood ratio (LLR)*

Convergence and Divergence was measured for each word level using the same procedure described in [9]. First, we tested each speaker-dependent model on the 60 Solo-pre words. Then we measured the posterior probability score for each word during interaction. We then set a threshold of 1.5 standard deviations (STD) based on the distribution of the prediction scores at baseline [as in 9]. If the word was predicted (i.e. minimum posterior probability) by the speaker-own model, we labeled that word as NoChange. If the word was predicted by its partner model we labeled that word as Convergent. Otherwise, when neither own or partner model predicted the word, we labeled that word as Divergent.

4. Results

4.1. Convergence and Divergence frequency

The total number of Convergence, Divergence and NoChange during the whole interaction is shown in Figure 3. Convergence and Divergence are variable phenomena because some dyads show a large amount of convergence while others much less [9][10]. The Female dyads (FF) converged more than male dyads (MM) (FF 25% and MM 7%) which is consistent with previous results [8][9]. A one-way repeated-measures ANOVA with the sessions (6 levels) as within-subject factor did not reveal any significant effect of Convergence ($F_{(5,45)}$ = 0.78, p=0.56). The same analysis on Divergence showed no significant effects ($F_{(5,45)}$ = 0.69, p=0.63) indicating that the amount of Convergence and Divergence did not change significantly across the experimental blocks.

4.2. Acoustic features

Four speech acoustic features, F0, F1, F2 and Intensity were extracted from the audio recordings using Praat software [11]. First, we averaged within each word and then within subjects. A two-tailed t-test, on z-scored values, was used to explore differences between Convergence and NoChange or Divergence and NoChange. Results show (Table 1) that intensity was significantly different ($t_{(9)}$= 4.93; p<0.0001) during Convergence compared to NoChange and during Divergence compared to NoChange ($t_{(9)}$= -2.81; p=0.02).

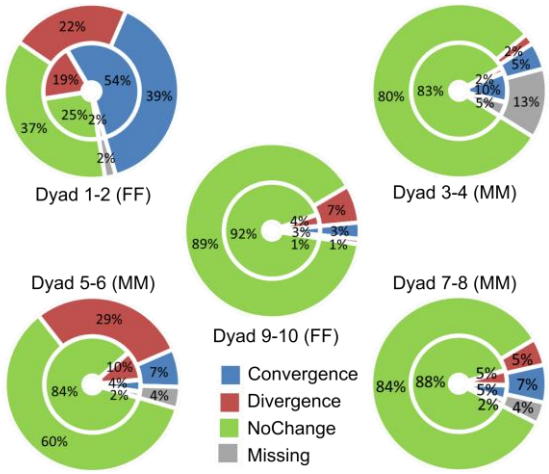

Figure 3. *No. of times each participant converged or diverged during the experiment.*

Table1: *Results of two-sided student t-test (Convergence Vs. NoChange and Divergence Vs. NoChange).*

	Conv (±STD)	**NoCh (±STD)**	**Div (±STD)**	**t-test p-value**	
				Con-NoCh	**Div-NoCh-**
F0 (Hz)	135±38	134±40	136±36	0.37	0.3
F1 (Hz)	83±22	69±28	80±24	0.56	0.09
F2 (Hz)	272±69	244±84	281±75	0.28	0.11
Intensity (dB)	58±10	58±10	57±10	**0.0001**	**0.02**

4.3. Vocal Tract features characteristics during accommodation

In the construction of our VDT list we included 11 different vowels (lexical code: $a, e, i, o, u, y, O, E, @, §, 2$; for details, see http://www.lexique.org/). However, given that for some subjects we observed relatively few instances of Convergence or Divergence we ended up with a smaller set of vowels in these categories. Therefore, to avoid any biased comparison, when analyzing articulatory data, we excluded NoChange words containing very rare vowels in Convergence or Divergence. A NoChange word was included if all its vowels were present in at least 5% of the convergent or divergent words. This resulted in four vowels ($/a/, /e/, /i/, /o/$) whose distribution in the three different conditions is shown in Figure 4. A two-tailed t-test on z-scored values was used to explore the differences between Convergence and NoChange or Divergence and NoChange for each vowel. Results show that only for $/e/$, there was a significant difference between Convergence and NoChange ($t_{(9)}$= 2.77; p=0.022). This means that the following analyses on articulatory data were run on all 4 vowels ($/a/, /e/, /i/, /o/$) as well on ($/a/, /i/, /o/$).

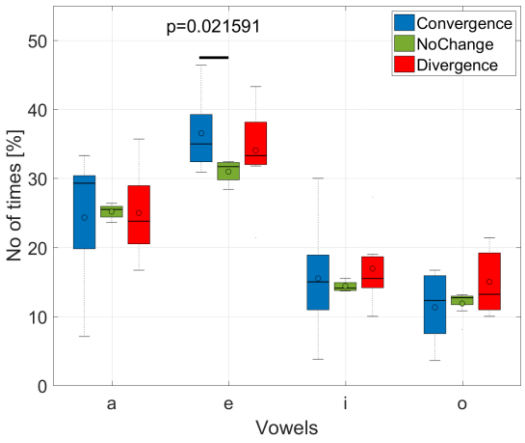

Figure 4: *Most frequent vowel distribution [%] in the three conditions for all subjects.*

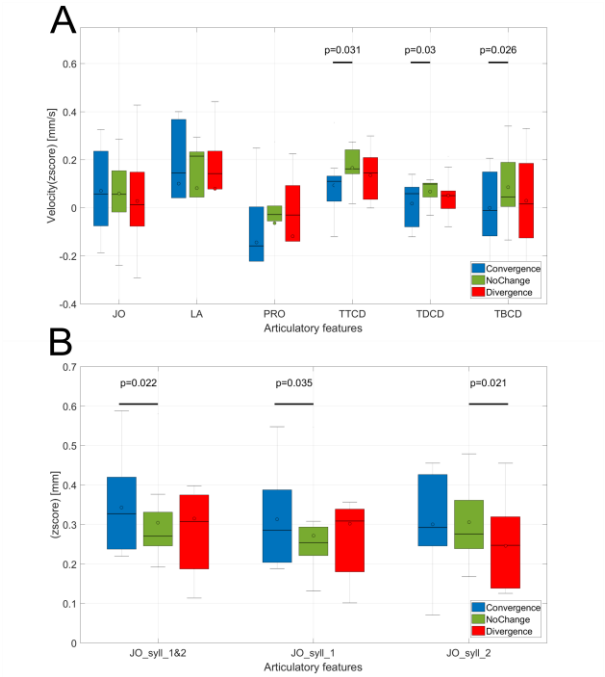

Figure 5*: (A) Whole-word articulatory data changes across conditions. The t-tests showed significant differences in the velocity of vocal tract tongue features (TTCD, TDCD, TBCD) between Convergence and NoChange (orizontal lines). (B) Syllable level differences in maximal jaw opening. JO_syll_1&2 and JO_syll_1 are significantly different in Convergence with respect to NoChange whereas JO_syll_2 is significantly different in Divergence with respect to NoChange.*

Vocal tract features were first averaged within each word then within each subject. A two-tailed t-test performed on z-scored values, was used to explore differences between Convergence and NoChange and between Divergence and NoChange. Results showed that velocity of the three vocal tract tongue features were significantly different in Convergence and NoChange conditions (TTCD: $t_{(9)}$= -2.55; p=0.031; TBCD: $t_{(7)}$= -2.82; p=0.026; TDCD: $t_{(8)}$= -2.63; p=0.03) (Figure 5A) demonstrating that during Convergence speakers move their tongue more slowly than in NoChange.

Moreover, maximal jaw opening was significantly modulated in Convergence Vs. NoChange (Figure 5B; JO_syll_1&2: $t_{(9)}$= 2.79; p=0.021; JO_syll_1: $t_{(9)}$= 2.47; p=0.03) and maximal jaw opening of the 2nd syllable was significantly different in Divergence Vs. NoChange (JO_syll_2: $t_{(9)}$= 2.75; p=0.022). Larger values in these features means that during Convergence speakers opened their jaw more than in NoChange, especially in the first syllable. Differently, in the second syllable the pattern of jaw opening was reversed and this was true for Divergence only. The same pattern is observed when removing the /e/ vowel from the dataset. Maximal jaw opening was significantly modulated in Convergence Vs. NoChange conditions (JO_syll_1&2: $t_{(9)}$= 3.13; p=0.012; JO_syll_1: $t_{(9)}$= 2.98; p=0.015) and maximal jaw opening of the 2nd syllable was significantly different in Divergence Vs. NoChange (JO_syll_2: $t_{(9)}$= 3.23; p=0.01).

5. Conclusion

Speech convergence is the phenomenon by which some participants in a dialogue tend to naturally align with each other in their phonetic characteristics. In this paper, we demonstrated the robustness of the automatic phonetic convergence detection method we already presented in [9]. In fact, as shown in Figures 2 and 3, our results were similar to those of our previous study. It is worth mentioning that the present dataset is characterized by relevant differences including participants' native language, the language of the word list, the word chain length, the pacing of VDT (self-paced as opposed to externally-paced) and the number of participants.

Besides, we also show preliminary but compelling results indicating that accommodation phenomena occur at the level of articulatory features. When we analyzed average velocity profiles at the whole-word level, we found that speakers slow-down their tongue movements during Convergence. Instead, when separating the two syllables of each word, we observed an interesting pattern of jaw maximal opening. In fact, the first syllable shows larger values during Convergence, whereas the second syllable smaller values for Divergence. Note that the first syllable is the one shared with the preceding word of the phonetic dyadic context (i.e. the word just uttered by the partner). Most importantly, the VDT rhyming rule forces subjects to focus their attention to the last syllable they heard to match it to the first they have to articulate. Interestingly, the opposite result we found for the second syllable could be explained by the fact that, for the speaker, this syllable does not have to comply with any specific rule. However, due to the variability of accommodation phenomena [10], results could in part be driven by dyads showing greater effects.

The present work starts exploring the articulatory counterpart of phonetic convergence. Future experiments will need to acquire larger number of dyads and eventually explore if at the single syllable level there are critical articulatory features [12] which are more or less robust to accommodation phenomena occurring during speech interactions.

6. Acknowledgement

We thank ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and the Excellence Initiative of Aix-Marseille University (A*MIDEX) for support.

7. References

[1] Schilbach, Leonhard, Bert Timmermans, Vasudevi Reddy, Alan Costall, Gary Bente, Tobias Schlicht, and Kai Vogeley. "Toward a second-person neuroscience 1." *Behavioral and brain sciences* 36, no. 4 (2013): 393-414.

[2] Giles, Howard. Communication accommodation theory. *John Wiley & Sons*, Inc., 2007.

[3] Goldinger, Stephen D. "Echoes of echoes? An episodic theory of lexical access." *Psychological review* 105.2 (1998): 251.

[4] Levitan, Rivka, and Julia Hirschberg. "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." *Interspeech*. 2011.

[5] Pardo, Jennifer S. "On phonetic convergence during conversational interaction." *The Journal of the Acoustical Society of America* 119.4 (2006): 2382-2393.

[6] Babel, Molly, and Dasha Bulatov. "The role of fundamental frequency in phonetic accommodation." *Language and Speech* 55.2 (2012): 231-248.

[7] Pardo, Jennifer S., Adelya Urmanche, Sherilyn Wilman, and Jaclyn Wiener. "Phonetic convergence across multiple measures and model talkers." *Attention, Perception, & Psychophysics* 79, no. 2 (2017): 637-659.

[8] Bailly, Gérard, and Amélie Martin. "Assessing objective characterizations of phonetic convergence." *Interspeech 2014*.

[9] Mukherjee, Sankar, Alessandro D'Ausilio, Noël Nguyen, Luciano Fadiga, and Leonardo Badino. "The Relationship Between F0 Synchrony and Speech Convergence in Dyadic Interaction." In *Interspeech 2017*, pp. 2341-2345. 2017.

[10] De Looze, Céline, et al. "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction." *Speech Communication* 58 (2014): 11-34.

[11] Boersma, Paul. "Praat: doing phonetics by computer." http://www. praat. org/ (2006).

[12] Stevens, Kenneth N. "On the quantal nature of speech." *Journal of phonetics* 17, no. 1 (1989): 3-45.