INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The Relationship between F0 Synchrony and Speech Convergence in Dyadic Interaction

Sankar Mukherjee[1], *Alessandro D'Ausilio*[1,3], *Noël Nguyen*[2], *Luciano Fadiga*[1,3], *Leonardo Badino*[1]

[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, Leonardo.Badino@iit.it

Abstract

Speech accommodation happens when two people engage in verbal conversation. In this paper two types of accommodation are investigated – one dependent on cognitive, physiological, functional and social constraints (Convergence), the other dependent on linguistic and paralinguistic factors (Synchrony). Convergence refers to the situation when two speakers' speech characteristics move towards a common point. Synchrony happens if speakers' prosodic features become correlated over time. Here we analyze relations between the two phenomena at the single word level. Although calculation of Synchrony is fairly straightforward, measuring Convergence is even more problematic as proved by a long history of debates on how to define it. In this paper we consider Convergence as an emergent behavior and investigate it by developing a robust and automatic method based on Gaussian Mixture Model (GMM). Our results show that high Synchrony of F0 between two speakers leads to greater amount of Convergence. This provides robust support for the idea that Synchrony and Convergence are interrelated processes, particularly in female participants.

Index Terms: Speech Convergence, Speech Synchrony, human-human interaction.

1. Introduction

When people engage in social interaction, they adjust their speech in order to accommodate to each other. This phenomenon is often labelled accommodation, imitation, convergence or alignment and synchronization. In this paper, we considered two types of accommodation – one due to cognitive, physiological, functional and social constraints [1], which we refer to as Convergence, and another due to linguistic and paralinguistic factors which we refer to as Synchrony [2][21][22]. Specifically, we consider that Convergence occurs when adjustments in both speakers' speech characteristics result in a shift towards a common point. Synchrony, on the other hand, refers to the situation where two speakers temporally display similar features, e.g., when one raises her/his voice intensity and the other speaker also raises her/his voice intensity.

We also define two other terms which are NoChange and Divergence. NoChange refers to the situation in which both speakers do not affect each other's behavior and their speech characteristics remain the same over the course of interaction. Divergence refers to the situation in which speakers move away from the speech of each other.

While the measurement of synchrony is straightforward, the quantification of speech Convergence is an open area of research. Previous research has dealt with objective acoustic measures [4] [5], while others have focused on subjective evaluations [6] [7]. Nevertheless, a great deal of inconsistency and variability exists among objective acoustic measurements [8]. Further complexity is driven by the temporal evolution of Convergence during interaction. Most of the research in this area has modeled Convergence as a linear process, i.e., it grows as the conversation proceeds [9] [10]. However, subjects do not remain involved at the same degree over the whole course of a conversation, suggesting that Convergence can be a time-varying phenomenon [2] [5] [11] [12].

In summary, Convergence is more likely to be both a linear and a dynamic phenomenon [5], can be achieved in multiple features (i.e. F0, intensity, etc.) and at multiple levels (phoneme, word, sentence, discourse). The aim of the present study is to find out the relation between Convergence and Synchrony.

2. Our Approach

To circumvent some of the problems that hamper an effective and robust measurement of Convergence, we did not use spontaneous conversations. Rather we used a constrained interaction task that allows better experimental control. Convergence is computed by using an automatic speaker identification technique to quantify subject's effort in moving towards the other speaker. Finally, we implemented a robust method to combine both participants' shift towards each other and explain their behavior over time.

To this purpose, i.e. to limit the complexity of the task while retaining the dynamic nature of true dyadic interaction, we developed a modified version of the Domino task [3], a controlled yet engaging speech interaction game. The Domino task consists in two speakers taking turn in chaining bi-syllabic words according to a rhyming rule: The first syllable of a word has to rhyme with the last syllable of the previous word. Differently from previous studies [3], we had native Italian speaker dyads do the Domino task in English as their second language (L2). Our hypothesis is that we will observe greater speech Convergence between interlocutors in an L2 – L2 interaction compared with an L1 – L1 interactions because there is more room for adjustment in the former case [13] [14] than in the latter case.

http://dx.doi.org/10.21437/Interspeech.2017-795

We did not force any hypothesis on what features to use, as we aimed to exploit the full richness of the acoustic spectrum by using Mel-frequency cepstral coefficients (MFCCs) [15]. Finally, a powerful, data driven, text independent, automatic speaker identification technique, based on GMM-UBM (Gaussian Mixture Model-Universal Background Model) was applied [3] to extract un-biased measures of Convergence. The Gaussian components model the underlying broad phonetic features that characterize a speaker's voice. In previous work [3] a similar method was applied except that the model was trained and tested on phonemes, whereas here we applied it on the whole word's acoustic form. This choice was motivated by the need to assess the relationship between Convergence and Synchrony. In fact, computing meaningful indexes of Synchrony requires that the features of interest (e.g., F0) be extracted from longer intervals than phoneme-sized ones.

3. Materials and method

3.1. Participants

For this experiment we recruited 16 native Italian speakers (8 males and 8 females, age: mean ± std; 26 years ± 2.3 years). Before the experiment, subjects were asked to self-rate their English knowledge on a 1-10 scale, including speaking fluency (7.19±1.17), reading (7.87±1.08), writing (7.31±0.95) and understanding (7.56±1.03). We grouped the subjects in 8 dyads (dyad 1 to 8), 4 female-female and 4 male-male. Before the start of the experiment subjects did not know each other and they did not interact with each other.

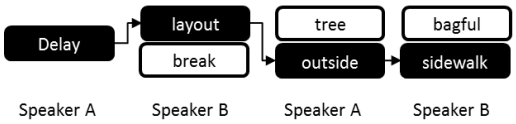

Figure 1. *Example of word domino*

3.2. Domino list

A verbal domino chain [3] was constructed with English words. To do this, we used the WebCelex (http://celex.mpi.nl/) English lexical database. Disyllabic words were first extracted from the database and then rearranged depending on spoken frequency (Collins Birmingham University International Language Database - COBUILD). A custom algorithm using R (https://github.com/sankar-mukherjee/SPIC-dommino) was then used to build the dominos. The algorithm starts from the highest lexical frequency word and then looks for the next highest frequency word, fulfilling the rhyming criteria and no repetitions. From the list generated, 200 unique bi-syllabic words were selected for the Verbal Domino task.

3.3. Procedure

The whole experiment was divided into three parts: Pre, Duet and Post. The verbal domino task was played on the Duet portion. 40 words were randomly selected from the 200-word chain. In Pre and Post, subjects had to read these 40 selected words individually. The Pre and Post parts were before and after the Duet respectively, and were used as baselines.

During the Pre and Post parts, subjects had to read aloud the 40 words presented on a screen one at a time. Between-word switching was controlled by a voice trigger. While one subject was performing this task, the other subject waited nearby. Each subject read 40 words in Pre and 40 in Post sections, for an overall 16x80 = 1280 words.

During the Duet part, the verbal domino task started with one word presented on the screen of one of the two subjects (say subject A) while the other partner (say subject B) was presented with a black screen. Then, when subject A read aloud that word, her/his screen immediately went black and subject B was presented with two words on her/his screen. When subject B read the word fulfilling the rhyming criteria, her/his screen went black and two words appeared on the screen of subject A, until the list ended (Figure 1). The voice onset triggered these changes. The whole experiment was monitored by one experimenter. In case of mistakes, subjects were told to stop and start again from the correct word.

We divided the Duet part in 4 sessions. The 200 selected words were divided into two 100-word chains. For the first two duet sessions, subject A initialized the chains, while for the other two duet sessions, subject B initialized the chains. Each subject read 50x4 = 200 words in the whole duet sessions. This resulted in a total of 3200 words. Only 98 errors out of 3200 words were recorded. The Duet task lasted about 25 minutes.

Between Pre, Duet and Post parts as well as between the 4 duet sessions, short breaks were introduced to allow the participants to rest.

The subjects' speech was recorded with a 44100 Hz sampling frequency – using two high-quality microphones (AKG C1000S) connected to an external dedicated audio mixer (M-Audio Fast Track USB II Audio Interface). An adaptive energy-based speech detector [16] was used for voice onset detection. All operations were implemented through a Psychotoolbox 3 script running in the Matlab environment.

4. Acoustic analysis

4.1. Pre-Processing

All words in which the voice trigger was incorrect (e.g. breathing, stuttering, etc., resulting in premature triggering or no triggering), or the word response was not correct, were removed for the analysis. This resulted in the removal of 98 out of 3200 words for the Duet and 33 out of 1280 for the Pre and Post parts. For each dyad, an average of 387.75±16.36 words were collected.

Periods of silence before and after each word were removed using an energy-based Voice Activity Detector. Then 39 dimension (13 static, 13 delta and 13 delta-delta) MFCCs (Mel Frequency Cepstral Coefficients) were extracted every 5ms from 10ms Hanning windows. Finally, MFCCs were z-score normalized to have 0 mean and 1 standard deviation to mitigate the effects of mismatch between microphones and recording environments.

4.2. GMM-UBM

The MSR Identity Toolbox [18] was used for GMM-UBM modelling. First the UBM was trained with the Pre data of all the subjects (consisting of a total of 124068 speech frames).

Then, individual speaker-dependent models were created using the Pre data of each speaker via maximum *a posteriori* (MAP) adaptation of the UBM (Figure 3A). The GMM-UBM has multiple hyper-parameters that can affect the speaker-dependent model performance. To check the quality of the speaker-dependent models, Post data (which had the same words as Pre) were used as a validation set. Each speaker-dependent model performance was verified against the UBM model. The confusion matrix for the cross validation set shows that modelling performance is fairly good (Equal error rate (EER) for the training: 2.26%, validation: 10.55%) as shown in Figure 2. After the test, we chose 32-component GMMs.

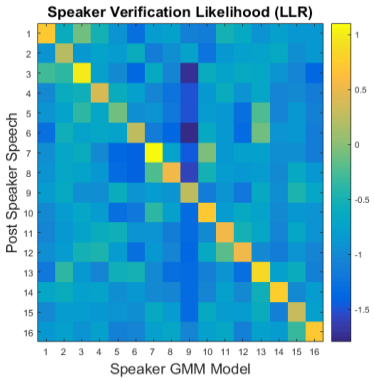

Figure 2. *Speaker verification confusion matrix of all the speaker-dependent models against background UBM in the Post data. Here the diagonal positive score line indicates a good MAP adaptation. The diagonal line (top left to bottom right) represents speaker dependent models performance on its their own speaker's speech (which is high compared to the others, suggesting a good speaker dependent model adaptation).*

Critical for our analysis is a speaker recognition decision. It mainly consists in a basic statistical test between two hypotheses:

$$\begin{cases} H_A: & \text{Speech y has been produced by the hypothesized speaker A} \\ H_B: & \text{Speech y has been produced by the hypothesized speaker B} \end{cases}$$

In our case, H_A and H_B are the models of the two speakers of the dyad. We then computed the log-likelihood ratio score (LLR) of samples y (here at single word level) using the following equation –

$$\mathrm{LLR(y)} = \log\left(\frac{p(y|H_A)}{p(y|H_B)}\right) \qquad (1)$$

Where H_A and H_B are the speaker models of speaker A and speaker B respectively. According to Eq.1, a positive LLR score means that the test speech is closer to speaker A than to speaker B, a negative value indicates an opposite pattern. A score close to zero means that the tested speech *y* has the same amount of probability of belonging to both classes. This means that the tested speech moved towards an average of both speakers' acoustic spaces, which is our definition of Convergence for one speaker.

4.3. Convergence calculation

In our experiment, Convergence was computed over time, i.e., if two consecutive LLR scores, one for each Duet participant, were close to each other, then they were regarded as acoustically more similar and phonetically convergent (Figure 3B). To verify whether LLR scores were signaling true Convergence/Divergence we created two criteria.

First, Pre data LLR scores of each subject were compared with their Duet data. Here, Pre data were used as a baseline. We computed how far the Duet LLR scores were from the Pre LLR score distribution. Each speaker Pre LLR score distribution was z-score transformed. Duet LLR scores which lay 1.5 std. away from the mean Pre LLR score distribution were considered as Convergence or Divergence points (see Figure 3B).

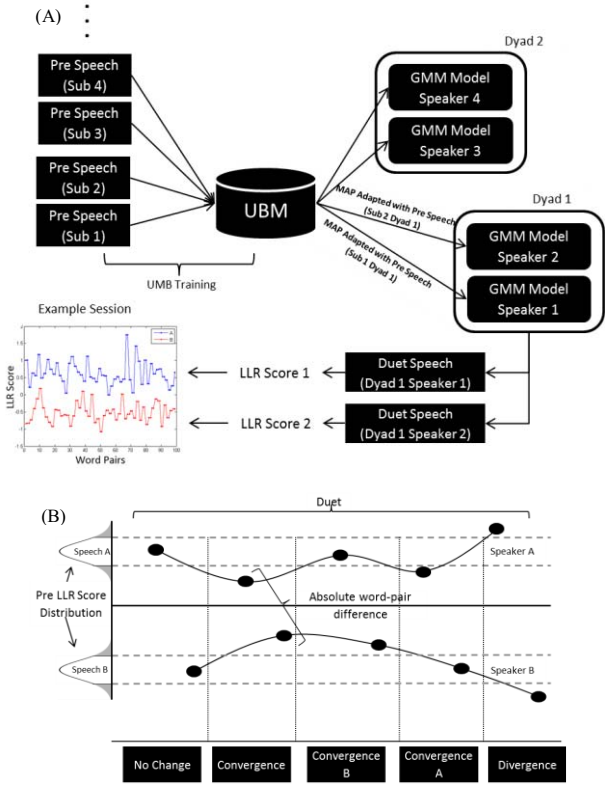

Figure 3. *(A) Schematic diagram of GMM-UBM modeling and how LLR score of test speech is predicted. (B) Graphical depiction of the strategy used to select word pair points. The strategy is – based on how the 2 points are far from the Pre distribution and how close they are to each other in the Duet condition.*

Second, we measured the absolute consecutive word dyad difference. To test whether these differences were meaningful and were not capturing accidental or coincidental phenomena [19] [20], true Duet data were compared to all possible pairings of participants who never actually interacted. This resulted in 48 surrogate Duets datasets (we did not combine male-female speakers). Absolute LLR score differences in two consecutive word pairs for both real and surrogate Duets were calculated. This surrogate distribution was z-score transformed. We considered it as background 'noise' and we

then controlled if absolute word-pair differences in true Duets lay 1.5 SD away from the mean.

Summing up, to be a Convergent one, a word pair had to meet two requirements: a) LLR score in the duet had to move towards the other speaker if compared to the distribution of speaker-specific LLR scores distribution in the Pre part; b) LLR score difference in the duet had to be a rare event if compared to the distribution of the same scores computed on random pairings of participants. In this paper we only considered Convergence and NoChange cases (See Figure 3B for a graphical depiction of the frequency of Convergence word pairs).

4.4. Synchrony measurement

To measure Synchrony four features were extracted: mean fundamental frequency (F0), mean duration, reaction time and mean intensity from each word. Praat software has been used to extract those features. Standard Pearson correlation coefficient $\rho_{xy} \in [-1,1]$ on two observation sets x and y (belonging to two separate subjects), was computed for each Duet session. This resulted in four correlation coefficients corresponding to each feature for each Duet session.

5. Results

5.1. Convergence results

After fulfilling the two Convergence criteria (see section on "Convergence calculation"), the number of Convergence points in each dyad was on average 12.62 % (std 9.02%). The total Convergence points of the whole experiment are shown in Figure 4 which indicates that Convergence is sparse. Some dyads had a large amount of convergence while others had a very limited one. Female dyads (FF) converged more than male dyads (MM) (FF 114 and MM 88) which is consistent with previous results [6] [3].

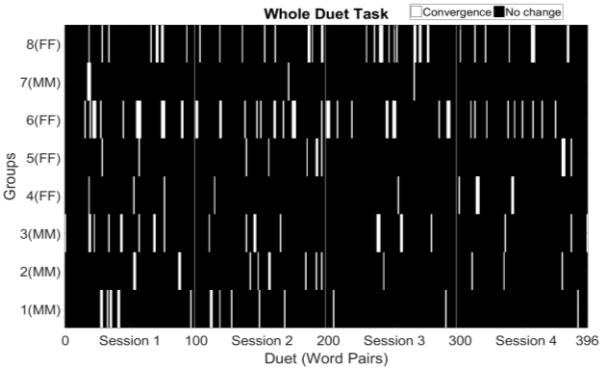

Figure 4. *No. of times when dyads converged in the whole experiment. White lines indicate the Convergence moments*.

5.2. Behavioral results

Differences in mean fundamental frequency (F0), mean duration, reaction time and mean intensity for each word during the Duet sessions between NoChange and Convergence trials are shown in Table 1. A two-sided Wilcoxon rank sum tests showed a significant difference in F0 and Intensity between Convergence and NoChange condition.

Table1: *Results of two-sided Wilcoxon rank sum test between Convergence and NoChange condition (mean±SD)*

	Convergence	**NoChange**	**Significance**
F0 (Hz)	136.71±46.58	127.85±43.98	$P<0.01$
Intensity (dB)	71.05±4.29	70.47±5.07	$P<0.05$
Duration (ms)	893±302	838±232	$P=0.11$
Reaction Time (ms)	426±298	427±262	$P=0.69$

5.3. Relation between Convergence and Synchronization

In order to establish the relationship between Synchrony and Convergence, we used Pearson correlation between the number of Convergence points and Synchrony correlation coefficient of each session. This resulted in a highly significant correlation for F0 (Table 2). This shows that Synchrony in F0 is associated to Convergence and this result was largely significant, for female-female dyads (Table 2).

Table2: *Correlation results between Synchrony and Convergence for males (MM) and females (FF) dyads*

Features	**All dyads**		**FF dyads**		**MM dyads**	
	CC	**sig (*P*)**	**CC**	**sig (*P*)**	**CC**	**sig (*P*)**
F0	**0.517**	**0.002**	**0.578**	**0.02**	0.199	0.459
Intensity	0.326	0.07	0.466	0.06	0.064	0.812
Duration	0.087	0.63	0.336	0.20	-0.293	0.269
Reaction Time	-0.002	0.99	-0.279	0.29	0.441	0.086

6. Conclusion

In this paper we show that speech Convergence can be measured using a speaker identification technique during a well constrained task such as the Domino [1] [3]. Importantly, we introduced several analysis features to make the estimation of Convergence more robust. For instance, we tested modelling performance and verified its validity. We also evaluated if Convergence scores were attributable to random fluctuations in the data or were the true effect of dyadic interaction by testing them against surrogate dyads. Results show that the nature of speech Convergence is sparse, i.e., it is not evenly distributed on all the dyads. Some dyads show higher degree of Convergence while others rarely converge at all. A possible factor in this sparseness may be due to subjects' attention, familiarity with the content and their likability towards each other. However small and sparse, Convergence was associated to Synchrony in F0. This is an interesting new addition to the current discussion about the nature of these two complementary aspects of speech accommodation. Our work provides support for the idea that Synchrony and Convergence are interrelated processes, particularly in female dyads.

Future work includes testing this speaker identification technique on free flow dialog and in L1-L1 settings.

7. References

[1] Littlejohn, Stephen W., and Karen A. Foss. Theories of human communication. Waveland press, 2010.

[2] Heldner, Mattias, and Jens Edlund. "Pauses, gaps and overlaps in conversations." Journal of Phonetics 38.4 (2010): 555-568.

[3] Bailly, Gérard, and Amélie Martin. "Assessing objective characterizations of phonetic convergence." *15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*.

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

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

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

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

[8] Pardo, Jennifer. "Measuring phonetic convergence in speech production." Frontiers in psychology 4 (2013): 559.

[9] Suzuki, Noriko, and Yasuhiro Katagiri. "Prosodic alignment in human–computer interaction." Connection Science 19.2 (2007): 131-141.

[10] Natale, Michael. "Convergence of mean vocal intensity in dyadic communication as a function of social desirability." Journal of Personality and Social Psychology 32.5 (1975): 790.

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

[12] Vaughan, Brian. "Prosodic synchrony in co-operative task-based dialogues: A measure of agreement and disagreement." (2011).

[13] Branigan, Holly P., Martin J. Pickering, and Alexandra A. Cleland. "Syntactic co-ordination in dialogue." Cognition 75.2 (2000): B13-B25.

[14] Kim, Midam, William S. Horton, and Ann R. Bradlow. "Phonetic convergence in spontaneous conversations as a function of interlocutor language distance." Laboratory phonology 2.1 (2011): 125-156.

[15] Aubanel, Vincent, and Noël Nguyen. "Automatic recognition of regional phonological variation in conversational interaction." Speech Communication 52.6 (2010): 577-586.

[16] Reynolds, Douglas A. A Gaussian mixture modeling approach to text-independent speaker identification. Diss. Georgia Institute of Technology, 1992.

[17] Pelecanos, Jason, and Sridha Sridharan. "Feature warping for robust speaker verification." (2001): 213-218.

[18] Sadjadi, Seyed Omid, Malcolm Slaney, and Larry Heck. "Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research." Speech and Language Processing Technical Committee Newsletter 1.4 (2013).

[19] Ramseyer, Fabian, and Wolfgang Tschacher. "Nonverbal synchrony or random coincidence? How to tell the difference." Development of multimodal interfaces: active listening and synchrony. Springer Berlin Heidelberg, 2010. 182-196.

[20] Ward, Arthur, and Diane J. Litman. "Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora." (2007).

[21] Rachel Coulston, Sharon Oviatt, and Courtney Darves. 2002. Amplitude convergence in children's conversational speech with animated personas. In Proc. ICSLP, volume 4, pages 2689–2692.

[22] Tatsuya Kawahara, Takashi Yamaguchi, Miki Uesato, Koichiro Yoshino, and Katsuya Takanashi. 2015. Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening. In Proc. APSIPA, pages 392–395. IEEE.