Studying the link between inter-speaker coordination and speech imitation through human-machine interactions

Leonardo Lancia, Thierry Chaminade, Noël Nguyen, Laurent Prevot

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

Studying the link between inter-speaker coordination and speech imitation through human-machine interactions

Leonardo Lancia[1], *Thierry Chaminade*[2], *Noël Nguyen*[3], *Laurent Prévot*[3]

[1] Laboratoire de Phonétique et Phonologie, France
[2] Institut de Neuroscience de la Timone, France
[3] Laboratoire Parole et Langage, France

leonardo.lancia@univ-paris3.fr,
thierry.chaminade@univ-amu.fr, noel.nguyen-trong@univ-amu.fr, laurent.prevot@univ-amu.fr

Abstract

According to accounts of inter-speaker coordination based on internal predictive models, speakers tend to imitate each other each time they need to coordinate their behavior. According to accounts based on the notion of dynamical coupling, imitation should be observed only if it helps stabilizing the specific coordinative pattern produced by the interlocutors or if it is a direct consequence of inter-speaker coordination. To compare these accounts, we implemented an artificial agent designed to repeat a speech utterance while coordinating its behavior with that of a human speaker performing the same task. We asked 10 Italian speakers to repeat the utterance /topkop/ simultaneously with the agent during short time intervals. In some interactions, the agent was parameterized to cooperate with the speakers (by producing its syllables simultaneously with those of the human) while in others it was parameterized to compete with them (by producing its syllables in-between those of the human). A positive correlation between the stability of inter-speaker coordination and the degree of f0 imitation was observed only in cooperative interactions. However, in line with accounts based on prediction, speakers imitate the f0 of the agent regardless of whether this is parameterized to cooperate or to compete with them.

Index Terms: verbal interactions, coordination, human-computer interaction, human dynamical clamp, prediction.

1. Introduction

Coordination between speakers is a crucial ingredient of speech communication. However, studying inter-individual coordination during verbal interactions is a hard task, ideally requiring real-time manipulation of the input to each speaker based on her/his behavior and on the behavior of her/his partner(s). The first aim of this contribution is to propose an experimental set-up allowing this kind of manipulation by adapting to speech the human dynamical clamp paradigm [1], originally proposed to study limb motor control. The second aim of this contribution is to test the predictions of two different ways of explaining how speakers coordinate during verbal interactions. Explanations based on internal predictive models [2, 3] propose that each individual predicts the behavior of the interlocutors by means of the forward models usually employed in controlling her/his own behavior. Explanations based on the notion of coupling between dynamical systems (henceforth dynamical coupling) [4, 5] propose that speakers do not need to predict the behavior of their interlocutors because coordinated behavior results from mutual dependencies between the dynamics governing the sensorimotor systems of the speakers engaging in a conversational interaction. In principle, both approaches can explain implicit imitation phenomena, often observed in speech (e.g. [6, 7]). According to explanations based on predictive models, imitation is observed because the internal forward models of the speakers are tuned to predict the behavior of their interlocutors. This account suggests that inter-speaker coordination always induces a tendency to imitate the interlocutor. An alternative account of imitation is based on the mutual constraints that link the moment-to-moment activities of the sensorimotor systems of speakers in interaction [4]. If coordination propagates across levels of activity, time scales and subsystems, it can induce behavioral matching by constraining the functioning of the individual sensory-motor systems. However, to the extent that imitation depends on coordination, it is expected to depend on the particular coordinative relation between the speakers and on its stability.

To test these predictions, we designed an artificial agent able to produce repeatedly a simple speech utterance while coordinating its behavior with a human speaker instructed to perform the same task. Pilot work showed that human speakers naturally tend to produce their syllables simultaneously with those of the agent (i.e. to coordinate their syllabic cycles in-phase with those of the agent). Moreover, when the agent is parameterized to produce its syllables in between those of the human (i.e. to coordinate its syllabic cycles in anti-phase with those of the human), speakers' fluency decreases and mispronunciations are observed more often. If imitation is due to the re-parameterization of the internal predictive models, we would expect to observe imitative behavior both when the agent is programmed to cooperate with the speaker (by targeting in-phase coordination) and when it competes with them (by targeting anti-phase coordination). If however imitation results from coupling between the speakers' and the agent's behaviors, the degree of imitation is expected to be correlated with the stability of the coordinative relation targeted by the speakers. This in turn is expected to be higher in cooperative interactions.

2. Design of the artificial agent

The artificial agent is a simplified speech synthesizer that repeats the utterance /topkop/ by varying its speech rate through WSOLA synthesis [8]. In this utterance, the alternation of two different tongue gestures at the onsets of otherwise identical syllables renders the articulation harder and harder as speech rate is increased [9]. This provides a mean to control the difficulty of the task for the human speaker. Moreover, due to the alternation of voiced nuclei with voiceless onsets and codas, it is easy to identify the syllabic cycles. To build the audio signal produced by the agent, we recorded a production of

http://dx.doi.org/10.21437/Interspeech.2017-1431

/topkop/ by a male Italian speaker (the prototype signal). At intervals of 12.5 ms the agent extracts and analyses the last portion of the input audio stream containing the speaker's speech signal (the input signal) and passes to the output audio stream a new portion of prototype signal (the output signal). The input and the output audio streams are updated through the Portaudio C library [10] accessed through the Psychophysics Toolbox for Matlab [11].

2.1. Main algorithm

The input signal at iteration n corresponds to the last chunk of the audio signal produced by the human digitized at 48000 Hz. Its duration is equal to the duration of three utterances produced at the expected speech rate. After decimation to 300 Hz, the input is rectified and low-pass filtered with cutoff frequency at 8 Hz. The obtained amplitude modulation signal (Am_{in}) captures acoustic energy variations mainly occurring at the time scale of syllables production. Am_{in} is submitted to a custom algorithm (see section 2.2) designed to estimate the instantaneous phase of its last value within the syllabic cycle produced by the speaker, denoted as $\phi_{in}(n)$. Likewise, $\phi_{out}(n)$ indicates the position inside the syllabic cycle of the last chunk of the prototype signal submitted to the output stream. The relative phase $\phi = \phi_{out}(n) - \phi_{in}(n)$, is submitted to a discrete-time variant of the Kuramoto equation [12].

$$\Delta\phi_{out}(n) = (r + k\,sin(\phi(n) + c))\Delta t \quad (1)$$

When the coupling parameter k, linking the behavior of the agent to that of the speaker, is equal to zero, the agent repeats the target utterance independently from the speaker at a rate determined by the constant r. When $k < 0$, the agent varies its speech rate from an iteration to the other in order to bring ϕ closer to c. When $k > 0$, the relative phase targeted by the agent is $\pi + c$. $|k|$ modulates the strength of the agent's tendency toward the target relative phase value and Δt is the time interval between each phase measurement and the next. In order to constrain the agent's speech rate, the position in the syllabic cycle reached by the agent at iteration n is determined by:

$$\phi_{out}(n) = min\left(max\left(\Delta\phi_{out}(n), \frac{r}{v}\right), rw\right) + \phi_{out}(n-1) \quad (2)$$

with r/v and rw determining respectively the minimum and the maximum speech rates. By transforming $\phi_{out}(n)$ from radians to samples of the prototype signal, we obtain the center sample of the next chunk of prototype signal to submit to the output stream. This location is corrected via the WSOLA algorithm [8] in order to avoid artefacts in the acoustic signal. Each output signal (duration: 25 ms) is combined to the output stream by overlap-add method (overlap: 12.5 ms).

2.2. Real-time measurement of instantaneous phase

The approach adopted to measure the instantaneous phase of Am_{in} is inspired by that proposed in [13] because it is suited for online processing. In this approach, the last state of the input signal is compared to a dictionary of states extracted from a recorded signal containing several repetitions of amplitude modulation cycles. Each state in the dictionary is associated to an instantaneous phase value. The instantaneous phase of the input state will be that of the dictionary state most similar to it.

The dictionary of states is built from amplitude modulation values extracted from 16 seconds of uninterrupted repetitions of the utterance /topkop/ produced by the same Italian speaker who produced the prototype signal. This signal (henceforth AM_{dic}), is submitted to time-delay embedding [14] allowing a better characterization of its cycles. This corresponds to building a m-dimensional time series $\overline{AM}_{dic}$ whose state at time i corresponds to a m-tuple of AM_{dic} values equi-spaced in time with lag τ:

$$\overline{AM}_{dic}(i) = AM_{dic}(i), AM_{dic}(i-\tau), AM_{dic}(i-2\tau), \ldots, AM_{dic}(i-m\tau) \qquad i = \{1, \ldots, N_{dic}\} \quad (3)$$

The dictionary of states is composed by the states of $\overline{AM}_{dic}$. Each state $\overline{AM}_{dic}(i)$ is associated to the instantaneous phase of $AM_{dic}(i)$. In the experiments described below, we used two prototype signals differing in f0 contours and we tested different speech rates. For each combination of speech rate and prototype f0 curve, we built a different dictionary of states by recording a sequence of utterances produced with the required f0 curve at the required speech rate.

For the algorithm to work properly, equivalent events occurring in the two cycles of amplitude modulation extracted from the prototype signal or in the cycles of amplitude modulation signal stored in the dictionary must be associated to a unique instantaneous phase value. To this end, we extracted the amplitude modulation cycles of the prototype signal and we created an average prototype cycle by time-aligning the two observed amplitude cycles to their average behavior via Functional Data Analysis registration [15] and by averaging again the two time-aligned cycles. Each sample of the average prototype cycle was assigned an instantaneous phase value starting at 0 and growing linearly toward 2π at the end of the cycle. By time-aligning the average prototype cycle, with the two cycles of the prototype signal and with the amplitude modulation cycles of AM_{dic}, we obtained a number of mapping functions (mapping each point of the average prototype cycle to a point in each of the aligned cycles). These were used to associate the instantaneous phase values from the average prototype cycle to the states of the amplitude modulation cycles in the prototype signal and in the dictionary.

To measure the instantaneous phase of the input signal, at each iteration of the main processing algorithm, Am_{in} is submitted to time delay embedding and the last state of the embedded time-series ($\overline{AM}^{*}_{in}$) is selected. In order to select the state of the dictionary corresponding to $\overline{AM}^{*}_{in}$, we computed:

$$D_n(i) = (\|\overline{AM}_{dic}(i), \overline{AM}^{*}_{in}(n)\|) + p\theta\big(R(i) - R(i_{n-1})\big) + q\theta\big(\phi_{dic}(i) - \phi_{dic}(i_{n-1})\big) \qquad i = \{1, \ldots, N_{dic}\} \quad (4)$$

Here $\|\cdot,\cdot\|$ is the Euclidean norm; $\overline{AM}_{dic}(i)$ is the i^{th} state in the dictionary; $\overline{AM}^{*}_{in}(n)$ is the last state of $\overline{AM}_{in}$ at iteration n; the parameter p modulates the strength of a first penalty term, favoring states belonging to the cycle containing the last selected state of the dictionary; θ is a step function ($\theta(x) = 1\ if\ x > 0, \theta(x) = 0\ if\ x \leq 0$); $R(i)$ is the integer associated to the syllabic cycle containing the i^{th} state of the dictionary; $R(i_{n-1})$ is the integer associated to the syllabic cycle containing the state of the dictionary selected at the previous iteration; the parameter q modulates the strength of a second penalty term, insuring that ϕ_{in} grows monotonically within each syllabic cycle; $\phi_{dic}(i)$ is the instantaneous phase associated to the i^{th} state in the dictionary; $\phi_{dic}(i_{n-1})$ is the instantaneous phase associated to the state of the dictionary selected at the previous iteration of the main algorithm. The instantaneous phase of the input signal at iteration n is $\phi_{in}(n) = \phi_{dic}\big(argmin(D_n)\big)$.

To provide the human speakers with an online visual feedback about their speech rate, at each iteration of the main processing algorithm AM_{in} is normalized and submitted to the

Hilbert transform. The obtained time-varying measure of instantaneous phase is then unwrapped. Speech rate is estimated by computing the mean difference between consecutive values. On a computer screen, we displayed a vertical green bar whose height varied over time indicating the speaker's current speech rate. A horizontal line at the appropriate height indicated the target speech rate (as determined by parameter r in eq. 1). To help the human speaker reaching that speech rate, during the first three repetitions of the prototype signal in each trial, the agent operated at the target speech rate without adapting to the speaker's behavior (i.e $k = 0$ in eq. 1).

3. Experiment

3.1. Procedure and participants

The experimental session started with a training composed minimally of three trials. In each trial, the speaker was asked to repeat without interruption during 13 sec the utterance /topkop/ at one of the speech rates used in the test section of the experiment. Speech rate increased from one training trial to the next and each training trial could be repeated at wish by the participant. In each trial of the test session, the human speaker was asked to repeat simultaneously with the agent the utterance /topkop/ during 13 secs without interruptions. Two different versions of the prototype signal were used. In half of the trials, the agent produced utterances with the prosodic prominence on their first syllables, in the other half the prominence was on the last syllables. Parameters varying across trials were k (vals.: -0.2, -0.15, -0.03, 0, 0.03, 0.15, 2) and r (vals.: 1.32, 1.764, 1.98 roughly corresponding to 1.7, 2.3 and 2.6 utt./sec.).

The test session was composed of 51 trials: one trial per combination of k and r. For each speech rate, we recorded three additional trials (not analyzed here) in which the agent either was silent or only produced the first three repetitions of one target utterance (with $k = 0$). Sequences were randomly ordered inside blocks of constant r, with r increasing from one block to the next. 10 speakers of Neapolitan Italian (5 females and 5 males) aged between 18 and 43 participated in the experiment. The audio signals produced by the agent and those produced by the human speakers were both recorded at 48 KHz. Fixed parameters were $c = 0,\ v = 5,\ w = 2,\ q = +\infty,\ p = \sigma_{dic}/100$ (where σ_{dic} is the standard deviation of the peak values in the amplitude modulation cycles of the dictionary of states). The embedding parameters m and τ were determined separately for each combination of prototype signal and speech rate by analyzing the relative AM_{dic} signals. τ was determined by adopting the mutual information criterion [16], while m was determined through the false nearest neighbors criterion [17].

3.2. Analyses

Both the agent's and human's audio signals were decimated to 300Hz, rectified and low pass filtered with cut off frequency of 8Hz. The relative phase between the syllabic cycles of the two partners was computed by submitting the obtained amplitude modulations to the Hilbert transform and by computing their time-varying difference. Utterances produced by the human speakers were segmented at the syllable level and portions of signals corresponding to utterances containing speech errors were removed. For each utterance produced by the human speakers, we computed the ratio between the peak f0 values in the two syllables. Since prominence is associated to higher values of f0 in the agent's utterances, the difference between the f0 ratio values observed during exposure to utterance-initial prominence and those observed during exposure to utterance-final prominence was used to determine the degree of phonetic convergence between the human speaker and the agent at the level of f0. To estimate the degree of coordination between the partners, we submitted their amplitude modulation signals to joint recurrence analysis [18] modified by following [19], to deal with strongly nonstationary signals. Via this technique, given two signals, we extracted a time-varying coordination index ($CI(t)$) by averaging the two conditional probabilities expressing the probability that the state observed at time t in one time series is repeated, given that the state observed at time t in the other time series is repeated. Repetitions of a state observed in a given syllabic cycle (either /top/ or /kop/) were retained only if found inside other tokens of that same syllabic cycle. We computed an utterance-specific mean coordination index ($UMCI$) by averaging the CI values observed in the time-interval corresponding to each utterance.

When $k > 0.03$, the relative phase was rarely smaller than $\pi/2$ (i.e. closer to in-phase). Therefore, to study the dependency between the degree of imitation and the stability of the in-phase coordination, statistical analyses were conducted only on utterances collected with $-0.03 \leq k \leq 0.03$. Since not all speakers were expected to show imitative behavior (see [20]), we excluded four speakers (three males and one female) who, when $k = 0$, did not show a significant degree of convergence toward the agent's f0 (for them f0 ratio did not change significantly when exposed to different f0 curves). A mixed model was run to determine the dependency of the f0 ratio on k, r, prominence location and on their double and triple interactions (number of observations: 1769). A second mixed model was run to determine the correlation between the f0 ratio and the index of inter-speaker coordination in those utterances where the relative phase was smaller than $\pi/2$ (number of observations: 1469). In this model, the predictors were the manipulated factors, the $UMCI$ and all their possible interactions. In each model, the random factors included speaker-specific and trial-specific intercepts as well as a speaker-specific slope for each predictor. Due to space limits, here we comment only results obtained at slow speech rate.

3.3. Results

Figure 1 displays three typical coordination patterns observed with different values of k. When the two partners cooperate ($k < 0$), they are coordinated in-phase. When they compete and the coupling is strong ($k = 0.15$), anti-phase coordination prevails. Finally, when the two partners compete with weak coupling (k=0.03), their relative phase wanders from 0 to 2π.

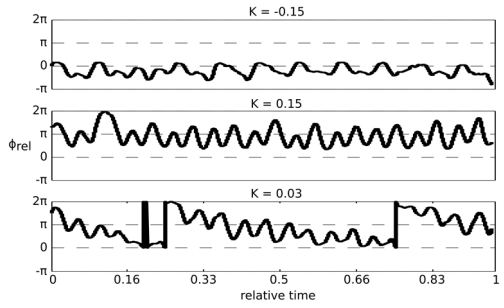

Figure 1: *Typical relative phase patterns obtained with different values of k.*

These patterns are reflected in the histograms of relative phase values obtained with different values of k (see Fig 2, left column). The right column of Figure 2 displays the boxplots of f0 ratio values over the prosodic make-up of the agent's utterances separately for different values of k. f0 ratios tend to be higher during exposure to utterance-initial prominence. When $k = 0$, the f0 ratio is significantly lower during exposure to utterance-final prominence ($estimate = -0.14, SE = 0.04, t = -3.4, p < 0.05$). No significant effect of k is observed, but the interaction between prominence location and k is significant and positive for $k = -0.03$ ($est. = 0.08, SE = 0.02, t = 4.37, p < 0.01$). This means that the difference due to the prosodic make-up of the stimuli is reduced when $k = -0.03$. This result suggests that speakers tend to imitate the f0 behavior of the agent across the board.

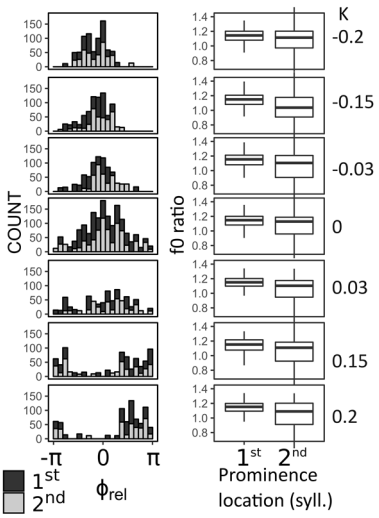

Figure 2: *Left column: histograms of mean relative phase values between the AMs of the two partners. Data are sorted by prominence location in the agent's utterances (black: 1st syllable, gray: 2nd syllable). Right column: box plots of f0 ratio values over prominence location (outliers not shown). In both columns, data are grouped in panels by value of k.*

Fig. 3 displays the predictions of the statistical model testing the correlation between the f0 ratio and the $UMCI$ (lines) superposed on the observed data (dots). At slow speech rate and with $k = 0$, f0 ratio and coordination index are negatively correlated during exposure to utterance-initial prominence (i.e. the effect of the $UMCI$ is negative: $est. = -0.02, SE = 0.01, t = -2.52, p < 0.05$). This correlation does not change significantly during exposure to utterance-final prominence (the effect of the interaction between the $UMCI$ and the prosodic make-up is not significant). However f0 is significantly reduced during exposure to utterance-final prominence ($est. = -0.12, SE = 0.05, t = -2.68, p < 0.05$). The difference in f0 ratio due to the stimuli prosody decreases when $k = -0.03$ (the interaction between $k = -0.03$ and the stimuli prosody is positive: $est. = 0.08, SE = 0.02, t = 3.95, p < 0.01$). The negative effect of the $UMCI$ observed during exposure to utterance-initial prominence is reduced when $k = -0.03$ (the interaction between $k = -0.03$ and the $UMCI$ is significantly positive: $est. = 0.03, SE = 0.01, t = 2.51, p < 0.05$). Moreover, the triple interaction between the $UMCI$, the stimuli prosody and $k = -0.03$ is significantly negative ($est. = -0.05, SE = 0.02, t = -3.46, p < 0.01$). Hence, when $k = -0.03$, the effects of $UMCI$ on f0 ratio diverge when speakers are exposed to different prosodic make-ups. This is interpreted as evidence that, when $k = -0.03$, the $UMCI$ and the degree of imitation are positively correlated. No significant difference is observed between $k = 0$ and $k = 0.03$.

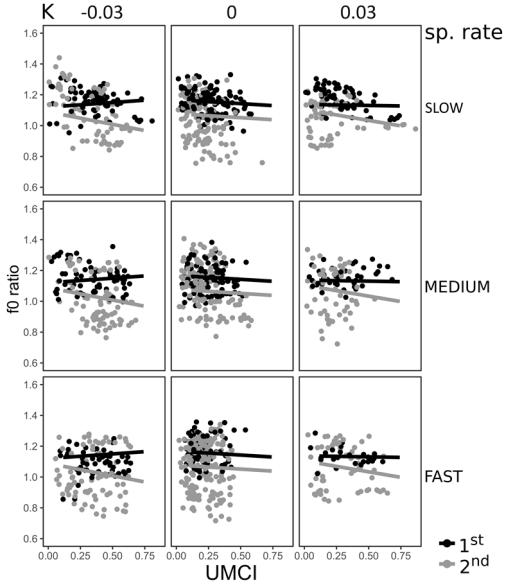

Figure 3: *f0 ratio over coordination index (UMCI). Lines represent the predictions of the mixed model. Data in each panel are grouped with respect to the location of prominence in the agent's utterances (1st or 2nd syll.). Only data included in the model are displayed.*

4. Discussion and conclusions

By adapting the human dynamical clamp paradigm to speech, we open the way to a new kind of experiments aimed at studying inter-speaker coordination by manipulating the unfolding over time of verbal interactions. The main goal of this study was understanding whether imitative tendencies in speech are due to the re-tuning of the internal predictive models supporting speech production or to dynamical coupling between the sensorimotor systems of the interlocutors. We hypothesized that if imitation of f0 curves depends exclusively on the dynamical coupling between the speakers it should be stronger in the cooperative condition (i.e. for negative k), because in this condition the coordination between the speakers is more stable (see the width of the distributions in the left column of Fig. 2). Since the degree of imitation does not change between competitive and cooperative interactions (see right column of Fig. 2), our results are more in line with accounts of inter-speaker coordination based on internal predictive models. However, the correlation observed between the degree of f0 imitation and the stability of inter-speaker coordination in the cooperative condition suggests that, given the appropriate conditions, inter-speaker coordination can spread through time scales and levels of activity and favor behavioral matching.

5. Acknowledgements

Research supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), ANR-10-LABX-0083 (EFL) and ANR-11-IDEX-0001-02 (A*MIDEX) of the French National Agency for Research (ANR).

6. References

[1] G. Dumas, G. C. de Guzman, E. Tognoli, and S. Kelso. "The human dynamic clamp as a paradigm for social interaction." *Proceedings of the National Academy of Sciences* vol. 111, no. 35 (2014): E3726-E3734.

[2] M. J. Pickering, and S. Garrod. "Forward models and their implications for production, comprehension, and dialogue." *Behavioral and Brain Sciences* vol. 36, no. 04 (2013): 377-392`.

[3] K. Friston, and C. Frith. "A duet for one." *Consciousness and cognition* vol. 36 (2015): 390-405.

[4] C. A. Fowler, M. J. Richardson, K. L. Marsh, and K. D. Shockley. "Language use, coordination, and the emergence of cooperative action." In *Coordination: Neural, behavioral and social dynamics*, pp. 261-279. Springer Berlin Heidelberg, 2008.

[5] D. Richardson, R. Dale, and K. Shockley. "Synchrony and swing in conversation: Coordination, temporal dynamics, and communication." *Embodied communication in humans and machines* (2008): 75-94.

[6] N. Nguyen, and V. Delvaux. "Role of imitation in the emergence of phonological systems." *Journal of Phonetics* vol. 53 (2015): 46-54.

[7] M. Sato, K. Grabski, M. Garnier, L. Granjon, J. L. Schwartz, and N. Nguyen. "Converging toward a common speech code: imitative and perceptuo-motor recalibration processes in speech production." *Frontiers in psychology* vol. 4 (2013): 422.

[8] W. Verhelst, and M. Roelands. "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech." In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2: 554-557. IEEE, 1993.

[9] L. Goldstein, M. Pouplier, L. Chen, E. Saltzman, and D. Byrd. "Dynamic action units slip in speech production errors." *Cognition* 103, no. 3 (2007): 386-412.

[10] Bencina, Ross, and Phil Burk. "PortAudio-an Open Source Cross Platform Audio API." In *ICMC*. 2001.

[11] M. Kleiner, D. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard. "What's new in Psychtoolbox-3." *Perception* vol. 36, no. 14 (2007): 1.

[12] Y. Kuramoto. *Chemical oscillations, waves, and turbulence*. Vol. 19. Springer Science & Business Media, 2012.

[13] A. Mörtl, T. Lorenz, and S. Hirche. "Rhythm patterns interaction-synchronization behavior for human-robot joint action." *PloS one* vol. 9, no. 4 (2014): e95195.

[14] J. Frank, S. Mannor, and D. Precup. "Activity and gait recognition with time-delay embeddings." In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (2010): 1581-1586.

[15] J. C. Lucero, K. G. Munhall, V. L. Gracco, and J. O. Ramsay. "On the registration of time and the patterning of speech movements." *Journal of Speech, Language, and Hearing Research* vol. 40, no. 5 (1997): 1111-1117.

[16] H_S. Kim, R. Eykholt, and J. D. Salas. "Nonlinear dynamics, delay times, and embedding windows." *Physica D: Nonlinear Phenomena* vol. 127, no. 1 (1999): 48-60.

[17] M. B. Kennel, R. Brown, and H. DI Abarbanel. "Determining embedding dimension for phase-space reconstruction using a geometrical construction." *Physical review A* vol. 45, no. 6 (1992): 3403.

[18] C. Bandt, A. Groth, N. Marwan, M. C. Romano, M. Thiel, M. Rosenblum, and J. Kurths. "Analysis of bivariate coupling by means of recurrence." In *Mathematical Methods in Signal Processing and Digital Image Analysis*, pp. 153-182. Springer Berlin Heidelberg, 2008.

[19] L. Lancia, D. Voigt, and G. Krasovitskiy. "Characterization of laryngealization as irregular vocal fold vibration and interaction with prosodic prominence." *Journal of Phonetics* vol. 54 (2016): 80-97

[20] C. L. Alan, C. Abrego-Collier, and M. Sonderegger. "Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits." *PloS one* vol. 8, no. 9 (2013): e74746.