Interspeech 2024
1-5 September 2024, Kos, Greece

Behavioral evidence for higher speech rate convergence following natural than artificial time altered speech

Jérémy Giroud[1], *Jessica Lei*[1], *Kirsty Phillips*[1], *Matthew H. Davis*[1]

[1]MRC Cognition and Brain Sciences Unit, University of Cambridge, UK

Jeremy.Giroud@mrc-cbu.cam.ac.uk, Matt.Davis@mrc-cbu.cam.ac.uk

Abstract

As AI progresses, our exposure to artificially generated spoken content varying in naturalness and speed increases. This trend is amplified by AI-powered personal assistants' proliferation, multiplying our interactions with intelligent systems. Research is crucial to understand if phenomena observed in human-human interactions can inform these new interactions. For instance, in everyday conversation, people adjust their speaking rate to match their partner's, a phenomenon known as speech rate convergence. It is crucial for effective communication, occurs automatically and is present in more artificial interaction scenarios. We investigated how the nature (natural vs. artificial) and the presentation rate (normal vs. fast) of the speech signal impact speech rate convergence. Data from 116 participants across two experiments reveal higher convergence towards naturally produced speech compared to artificially time altered speech. Implications for human-machine interactions are discussed.

Index Terms: speech rate convergence, speech production, artificial natural speech

1. Introduction

1.1. Convergence

During conversations, interlocutors often non consciously alter their behavior to become more similar to their partner. This phenomenon is referred to as convergence. Evidence have been found for such convergence at different acoustic and linguistic levels (vocal intensity [1], speech rate [2], phonetic features [3], lexical features [4], syntactic structures [5]). Convergence is clearly seen in human-human interactions but is also reported for human-computer interactions [6, 7, 8]. In this work we explore the influence of the naturalness (or otherwise) of speech stimuli on convergence behavior; using speech rate convergence as a test case.

Speech rate convergence is a particular instance of convergence in which interlocutors tend to adopt a similar speech rate during conversations to achieve smooth inter-turn transitions and optimal comprehension. Speech rate is typically measured in syllables per second and is of particular interest because it reflects rhythmic information at the syllabic time scale which plays a crucial role for comprehension [9, 10, 11]. Moreover, speech rate changes modulate lower-level acoustic-phonetic and higher-level linguistic processing in parallel; for example, acoustic cues to segments change with speech rate (such as voice-onset time for stops; [12]) and change the processing time available for word identification, semantic access and syntactic integration [13]. Despite significant differences in speech rate across languages, speaking conditions, age and gender [14], listeners are remarkably good at adapting to even extreme variations in speech rate [15]. Moreover, there is now evidence that speech rate convergence takes place in human-human conversations [16]; with participants altering their speech rate to match the speech rate of confederates during dialogues [17]. Furthermore, listeners also adjust their speech rate to match audio recordings played at various rates demonstrating that this ability is not only restricted to human-human interaction [18].

1.2. Properties of naturally and artificially time altered speech signal

When interlocutors increase their speech rate, it induces multiple changes to the relative timing of speech units at the syllable level [19]: the length of between-word pauses are shortened, and the duration of vowels and unstressed syllables tend to be reduced more than the duration of consonants and stressed syllables [20]. As a result, naturally fast produced speech is accompanied by specific spectro-temporal changes that result in a range of nonlinear modifications to the speech signal [19]. In contrast, artificially time-altered speech results from linear modifications of the initial speech signal; that is consistently changing the spectro-temporal structure of all aspects of the original signal equally. Due to these differences in their acoustic properties, these two types of modifications may have different consequences at the behavioral and perceptual levels. Artificially time altered speech is easier to process compared to natural fast speech. Janse et al. [19] found that at 1.5 times normal rate, naturally produced fast words were less intelligible than linearly compressed words and required more time to process . In addition, under the same speech duration, natural fast speech (produced at 1.4 times normal speed) is also perceived as faster than linearly compressed speech [21].

1.3. Current study

The current study aimed at determining whether speech rate convergence differs between artificially and naturally produced time-altered speech. More specifically, in two online behavioral experiments, we (1) confirmed the effectiveness of audio recordings in inducing speech rate convergence, (2) determined whether participants exhibit comparable levels of speech rate convergence towards natural versus artificially time altered speech, (3) examined whether lasting effects on the rate of speech production emerge after a convergence period, and finally, (4) investigated whether prior familiarization with the type and rate of speech (during a perceptual adaptation period), influences speech rate convergence. A schematic of this procedure is shown in Figure 1a.

10.21437/Interspeech.2024-1001

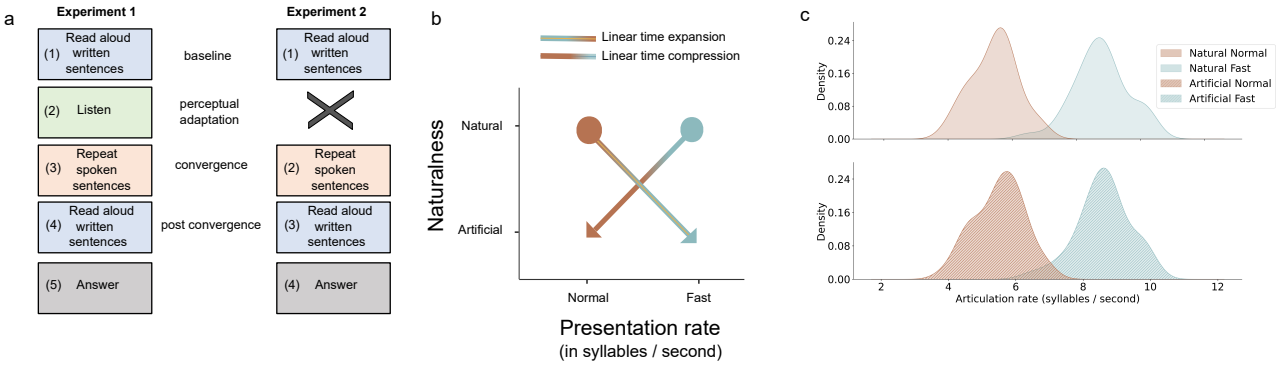

Figure 1: *Experimental procedure and stimuli creation. a) The study follows a block design where each block consists of 5 different tasks to be performed in the same order in Experiment 1 (left) and only 4 tasks in Experiment 2 (right). (1) Read aloud written sentences which allows to record participants' individual baseline articulation rate. (2) Listen enables participants to perceptually adapt to the presentation rate and naturalness of the speech signal in the current block through the listening of a short fable. (3) Repeat spoken sentences is used to assess participants' rate convergence to the audio recordings. (4) Read aloud written sentences records participants' individual articulation rate post convergence (after-effect). (5) Answer is an attentional check to assess participants' attention to the short fable. Participants completed all four blocks in a randomized manner. In Experiment 2, the Listen task was removed (as indicated by the grey cross) and participants were not able to perceptually adapt to speech characteristics. b) Natural normal speech (top left) is linearly compressed to create artificial fast speech (see red to blue arrow); similarly, natural fast speech (top right) is linearly expanded to create artificial normal speech (see blue to red arrow). The resulting stimuli have the same articulation rates while being of different nature (artificial/natural). c) Distributions of sentences' articulation rates for each experimental condition. Top: Natural conditions, bottom: Artificial conditions.*

2. Methods

2.1. Stimuli creation

We used an existing audio database, the CHAINS corpus, to create the speech stimuli [22]. The corpus contains 33 English sentences and 4 short stories (fables) produced at both normal and fast (54% faster) speed by different speakers. Recordings from one male speaker was used for the study. Linearly time-altered stimuli were created using the PSOLA algorithm [23]. Artificial fast speech was created through linearly compressing the natural normal stimuli so that the modified articulation rate of stimuli matches that of the corresponding natural fast stimuli, and similarly, artificial normal speech was created through linearly expanding the natural fast stimuli. The resulting stimuli had the same articulation rates while being of different nature (artificial / natural) (Figure 1.b).

2.2. Experimental procedure

Experiment 1 consisted of four experimental blocks (two by two design) in a randomized order (Figure 1.a). Each block contained 5 tasks performed always in the same order: (1) Participants first read aloud 6 written sentences presented on the screen (mean: 8 words), to get a baseline measurement of their speech rate. (2) They then listened to a short story (of approximately 39.5 seconds) to allow for perceptual adaptation to both the presentation rate and naturalness of the speech in each condition. (3) They then repeat 8 spoken sentences (mean: 9 words) which enabled us to assess convergence between their speech rate and that of the audio recordings. (4) Participants then read aloud written sentences a second time, identical to the first one, to measure their articulation rate following exposure to speech in the different experimental conditions (5) At the end of each block, participants answered two simple multiple-choice questions regarding the content of the short story to verify that they listened to the stories attentively. Experiment 2 consisted of the same set of experimental tasks, but the adaptation period (short story listening task) was removed. This design allowed us to assess the effect of this perceptual adaptation period on speech rate convergence.

2.3. Subjects and data acquisition

The two online experiments were created and hosted using jsPsych [24] and JATOS [25]. 61 and 55 subjects took part in Experiment 1 and Experiment 2, respectively. They were recruited via the online experimental platform Prolific and were financially compensated for their participation. Individuals who self-reported as having speech or language impairment or hearing problems were not eligible to participate. All subjects needed to pass a microphone test and a simple hearing test before they entered the online study. Before completing all experimental blocks containing the five different tasks, participants were presented with a short training session to get familiarized with the procedure. The experimental session lasted approximately 40 minutes in total. Informed consent was acquired directly through Prolific. Ethical approval was obtained from Cambridge Psychology Research Ethics Committee (CPREC).

2.4. Data analysis

Participants' audio responses were first subject to forced-alignment using the WebMAUS online service [26] and the output used to retrieve individuals' articulation rate for each condition. Articulation rate was computed as the number of syllables produced divided by the total duration of the utterance minus the marked duration of silent pauses. Articulation rate data were entered into a linear mixed effects analysis [27] to assess the effect of the experimental conditions on the speech rate convergence. The model included as fixed effects the two

condition factors (stimulus presentation rate – fast/normal; and naturalness (i.e. without or with linear temporal modifications; hereafter natural/artificial) and their interaction. In addition, we included participants' individual baseline articulation rate as a co-variate and also included by-subject and by-item random intercept. We focused our analyses on tasks in which articulation rate is measured ((3) repeat spoken sentences and (4) read aloud written sentences), enabling us to examine both speech rate convergence and after-effects.

3. Results

116 participants performed the speech perception and production tasks shown in Figure 1.a. They were presented with written transcriptions or audio recordings of spoken sentences and prompted to say them aloud.

3.1. Experiment 1

In Experiment 1, sixty one participants (25 females) performed the online experiment. A first linear mixed effects model analysis was conducted to examine whether and how participants' articulation rate was modulated by the characteristics of the audio recordings during the convergence period ((3) repeat spoken sentences). The model included participants and items (individual sentence) as random effects, speech naturalness (natural vs artificial), stimulus presentation rate (normal vs fast), their interaction and participants' baseline articulation rate as fixed effects. The full model accounted for 73% of the variance of the data and revealed a significant interaction of speech naturalness and speed ($\beta = 0.25 \pm 0.04, p < 0.001$), and a significant effect of individuals' baseline articulation rate ($\beta = 0.67 \pm 0.07, p < 0.001$). Bonferroni adjusted Post Hoc comparisons revealed that participant's articulation rate was significantly higher for natural stimuli at a fast presentation rate compared to natural stimuli at normal presentation rate ($p < 0.001$), and this was also the case for artificial stimuli at a fast presentation rate compared to artificial stimuli at normal presentation rate ($p < 0.001$). Between the normal and fast presentation rates, participants increase their articulation rate by 8% in the natural condition, while this increased is lower in the artificial condition with 3% increase in articulation rate. The results indicate that during the convergence task, participants display speech rate convergence towards audio recordings and this phenomenon is dependent on the specific characteristics of the speech signal.

The paradigm also allowed us to examine whether the convergence period elicits a sustained effect on participants' articulation rate. To investigate this issue, we built another linear mixed effects model to model participants' articulation rate during (4) read aloud written sentences. It included the same predictors as in the previous model. It accounted for 53% of the variance and revealed significant effects of stimulus presentation rate ($\beta = 0.14 \pm 0.05, p = 0.007$) and participants' baseline speech rate ($\beta = 0.53 \pm 0.06, p < 0.001$) only, indicating that participants showed sustained after-effects – i.e. changes in articulation rate during reading that depends on the rate of speech that was heard and repeated during the convergence task. However, unlike the speech rate convergence effects shown previously, these after-effects do not significantly differ in magnitude as a function of naturalness ($\beta = 0.02 \pm 0.05, p = 0.626$) .

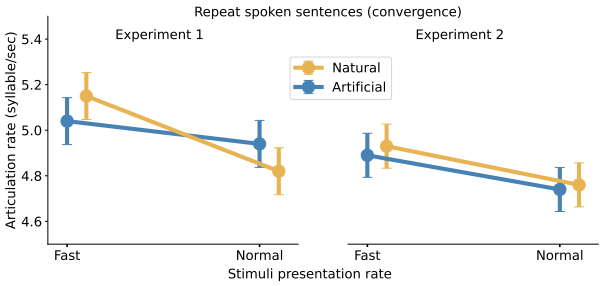

Figure 2: *Speech rate convergence from both experiments. Articulation rate data were extracted during the repeat spoken sentences task from n = 61 and n = 55 participants. Error bar represents standard error to the mean.*

3.2. Experiment 2

In Experiment 2, fifty five new participants (35 females) took part in a shorter version of the previous experiment, which did not include shorts stories (Figure 1.a, right). It was designed to assess the effect of the removal of a perceptual adaptation period (short story) on participants' speech rate convergence. As previously, we looked at both the convergence and the after-effect periods and modeled participants' articulation rate with linear mixed effects models. Analysis on data from the convergence period revealed that only stimulus presentation rate ($\beta = 0.25 \pm 0.04, p < 0.001$) and participants' baseline articulation rate ($\beta = 0.67 \pm 0.07, p < 0.001$) have a significant effect on speech rate convergence. There was no main effect or interaction with naturalness ($\beta = 0.03 \pm 0.02, p = 0.102$; $\beta = 0.02 \pm 0.04, p = 0.593$). The results suggest a limited convergence phenomenon during the convergence task in this second experiment. To look at potential after-effects we ran a linear mixed effects model on the articulation rate data recorded during the read aloud written sentences period. The model accounted for 59% of the variance and only revealed a significant effect of participants' baseline articulation rate ($\beta = 0.58 \pm 0.07, p < 0.001$), demonstrating no overall lasting effect following the convergence task in this second experiment.

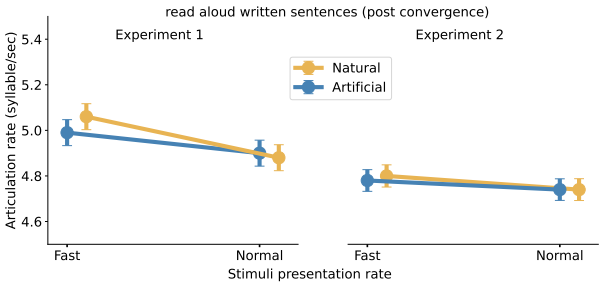

Figure 3: *post convergence after effects from both experiments. Articulation rate data were extracted during the read aloud written sentences task from n = 61 and n = 55 participants. Error bar represents standard error to the mean.*

Finally, to evaluate the modulation of the speech rate convergence phenomenon by experimental conditions across both experiments, we ran a final linear mixed effects model including articulation rate data obtained during the repeat spoken sentences period from the two experiments and added experiment as a fixed factor in the model. This final and larger

model explained 75% of the variance and revealed a significant effect of the three-way interaction (naturalness*presentation rate*experiment) ($\beta = 0.22 \pm 0.06, p < 0.001$), indicating that speech rate convergence is modulated by the characteristics of the speech signal as well as participants' familiarity with these characteristics (perceptual adaptation). Specifically, additional speech rate convergence is observed when: (1) participants listen to naturally-fast speech stimuli, and (2) when they have an extended period of listening to naturally fast-speech. The same inter experiments analysis was carried out on the after-effects period (read aloud written sentences period) from both experiments. The model did not reveal any statistically significant two-way or three-way interactions with naturalness, stimuli presentation rate and experiment during the post convergence period ($\beta = -0.05 \pm 0.06, p = 0.434$; $\beta = 0.06 \pm 0.13, p = 0.632$).

4. Discussion

In two behavioral experiments, we examined speech rate convergence for both naturally and artificially time altered speech using sentence reading and repetition tasks. Our results indicate that these tasks are sufficient to elicit speech rate convergence. Moreover, the magnitude of the convergence is dependent on the characteristics of the presentation rate and naturalness of heard speech. We also found evidence for modulation of participants' articulation rate beyond the initial convergence period; i.e. sustained speech-rate after-effects. Finally, convergence depends on participants' familiarity with the specific characteristics of the speech signal acquired during a perceptual adaptation period.

In our experiments, participants' repetition behavior varied systematically according to the acoustic characteristics of the audio recordings. This result is in line with previous work [28, 29, 17]. Although the increase in articulation rate is subtle (on average 5.8% while our stimulus manipulation is 59%), this is comparable to the effect size reported in other studies using real-life conversations [16, 17].Additionally, the current study implies that the phenomenon of convergence at play during human-human communication is not only restricted to naturalistic conversational interactions but also present in more experimentally controlled and/or artificial scenarios.

Moreover, we showed that speech rate convergence is more prominent when the rate of speech varied naturally compared to conditions in which stimulus presentation rate was artificially modified by linear time-compression or expansion. This is consistent with recent work showing that the magnitude of convergence is dependent on the nature of the speech signal and interlocutor; for example, reduced convergence towards artificially generated speech (computer voices from AI powered voiced assistants) compared to human produced speech [30].However, in the current study, our more artificial condition was created by linear modification of natural speech. Our data joins previous data in suggesting that artificially-modified speech is perceived as qualitatively different from naturally produced speech and that this affects vocal behavior. For instance, previous research has shown that linearly time compressed speech is easier to process and perceived as slower than natural fast speech [21]. This may explain the reduced tendency of participants to match the articulation rate of the stimuli; their perception of the speed of the stimuli may not accurately reflect its actual rate, thereby influencing their convergence behavior. In contrast, natural fast speech contains complex changes to spectro-temporal properties and is perceived as faster than artificial stimuli. In our study, this might have causing participants to increase their articulation rate to a greater degree; reinforcing the difference in the magnitude of the convergence between normal and fast speech conditions.

Our paradigm also allowed us to investigate how participants' vocal production characteristics change over an extended period outside of the convergence task. There are two results in our study that suggest longer-term adaptation. Firstly, comparing participants' baseline articulation rate with their articulation rate following convergence in Experiment 1 showed a small, but reliable after-effect. Changes in participants' articulation rate were seen following exposure to fast speech. This finding shows long-lasting changes in articulation rate remain apparent during reading aloud after speech rate convergence. In contrast, however, reliable after-effects were not observed in Experiment 2, despite reliable (though reduced) convergence effects. The primary difference between Experiment 1 and Experiment 2 is the exclusion of the story listening task. This design therefore enables to show the effect of a perceptual adaptation period on speech rate convergence and after-effects. In Experiment 2, participants had no opportunity to familiarize themselves with the specific acoustic characteristics of the speech signal prior to the convergence task. This resulted in an overall lower magnitude of speech rate convergence and abolished the after-effects that are apparent in Experiment 1. Thus, Experiment 2 shows the pivotal role of participants' familiarity with speech acoustic properties in eliciting long-lasting convergence.

Despite the interest and importance of our findings some limitations remain. One limitation pertains to ecological validity; our use of a sentence repetition task allows for good experimental control but might not be representative of convergence between interlocutors since conversations do not involve immediate, and complete repetition of heard speech. Alternative test tasks, such as structured or scripted turn-taking conversations could be used to increase ecological validity while retaining experimental control. A second limitation is that all participants heard the same, male model speaker. It has previously been found that male-male or male-female pairs produce greater speech rate convergence than female-female pairs [28, 17]. Future work can examine whether gender differences are apparent in convergence towards linearly altered or naturally produced fast speech.

Overall, our findings contribute to the broader scientific understanding of speech convergence. Our results suggest that speakers apply principles from human-human interaction while engaging with artificially generated speech. Critically, our data shows that participants interact in distinct ways depending on the rate and naturalness of the speech they are hearing. The findings from this study also have real-life implications for the design of voice-controlled human-machine interfaces and AI assistants. This study suggests that mimicking the spectro-temporal features of natural speech may yield better engagement from individuals and more user-friendly technologies than simpler, linear speech rate manipulations. In planned work we will explore which spectro-temporal features of story and sentence stimuli suffice to produce speech-rate convergence and/or after-effects, and how these cues relate to intelligibility.

5. References

[1] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability." *Journal of Personality and Social Psychology*, vol. 32, no. 5, p. 790, 1975.

[2] D. Freud, R. Ezrati-Vinacour, and O. Amir, "Speech rate adjust-

ment of adults during conversation," *Journal of fluency disorders*, vol. 57, pp. 1–10, 2018.

[3] M. Kim, W. S. Horton, and A. R. Bradlow, "Phonetic convergence in spontaneous conversations as a function of interlocutor language distance," 2011.

[4] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation." *Journal of experimental psychology: Learning, memory, and cognition*, vol. 22, no. 6, p. 1482, 1996.

[5] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic coordination in dialogue," *Cognition*, vol. 75, no. 2, pp. B13–B25, 2000.

[6] M. Cohn, B. F. Segedin, and G. Zellou, "Imitating siri: Socially-mediated vocal alignment to device and human voices," *Proc. 19th Int. Congr. Phon. Sci*, pp. 1813–1817, 2019.

[7] G. Zellou, M. Cohn, and B. Ferenc Segedin, "Age-and gender-related differences in speech alignment toward humans and voice-ai," *Frontiers in Communication*, vol. 5, p. 600361, 2021.

[8] G. Zellou, M. Cohn, and T. Kline, "The influence of conversational role on phonetic alignment toward voice-ai and human interlocutors," *Language, Cognition and Neuroscience*, vol. 36, no. 10, pp. 1298–1312, 2021.

[9] J. E. Peelle and M. H. Davis, "Neural oscillations carry speech rhythm through to comprehension," *Frontiers in psychology*, vol. 3, p. 320, 2012.

[10] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature neuroscience*, vol. 15, no. 4, pp. 511–517, 2012.

[11] J. Giroud, J. P. Lerousseau, F. Pellegrino, and B. Morillon, "The channel capacity of multilevel linguistic features constrains speech comprehension," *Cognition*, vol. 232, p. 105345, 2023.

[12] Q. Summerfield, "Articulatory rate and perceptual constancy in phonetic perception." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 7, no. 5, p. 1074, 1981.

[13] M. H. Christiansen and N. Chater, "The now-or-never bottleneck: A fundamental constraint on language," *Behavioral and brain sciences*, vol. 39, p. e62, 2016.

[14] D. Poeppel and M. F. Assaneo, "Speech rhythms and their neural foundations," *Nature reviews neuroscience*, vol. 21, no. 6, pp. 322–334, 2020.

[15] E. Dupoux and K. Green, "Perceptual adjustment to highly compressed speech: effects of talker and rate changes." *Journal of Experimental Psychology: Human perception and performance*, vol. 23, no. 3, p. 914, 1997.

[16] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evolution and Human Behavior*, vol. 34, no. 6, pp. 419–426, 2013.

[17] B. G. Schultz, I. O'BRIEN, N. Phillips, D. H. McFARLAND, D. Titone, and C. Palmer, "Speech rates converge in scripted turn-taking conversations," *Applied Psycholinguistics*, vol. 37, no. 5, pp. 1201–1220, 2016.

[18] M. K. Jungers, C. Palmer, and S. R. Speer, "Time after time: The coordinating influence of tempo in music and speech," *Cognitive Processing*, vol. 1, no. 2, pp. 21–35, 2002.

[19] E. Janse, "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech," *Speech communication*, vol. 42, no. 2, pp. 155–173, 2004.

[20] L. Max and A. J. Caruso, "Acoustic measures of temporal intervals across speaking rates: Variability of syllable-and phrase-level relative timing," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 5, pp. 1097–1100, 1997.

[21] E. Reinisch, "Natural fast speech is perceived as faster than linearly time-compressed speech," *Attention, Perception, & Psychophysics*, vol. 78, pp. 1203–1217, 2016.

[22] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains speech corpus: Characterizing individual speakers," in *Proc of SPECOM*, 2006, pp. 1–6.

[23] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[24] J. R. De Leeuw, "jspsych: A javascript library for creating behavioral experiments in a web browser," *Behavior research methods*, vol. 47, pp. 1–12, 2015.

[25] K. Lange, S. Kühn, and E. Filevich, "" just another tool for online studies"(jatos): An easy solution for setup and management of web servers supporting online studies," *PloS one*, vol. 10, no. 6, p. e0130834, 2015.

[26] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[27] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[28] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[29] J. S. Pardo, I. C. Jay, and R. M. Krauss, "Conversational role influences speech imitation," *Attention, Perception, & Psychophysics*, vol. 72, no. 8, pp. 2254–2264, 2010.

[30] G. Zellou and M. Cohn, "Social and functional pressures in vocal alignment: Differences for human and voice-ai interlocutors," in *Proceedings of Interspeech*, 2020.