The One-Voice Expert

Lisa Evelyn Rombout (l.e.rombout@uvt.nl)
Department of Cognitive Science and Artificial Intelligence, Tilburg University,
Warandelaan 2, 5037 AB Tilburg, The Netherlands

Marie Postma-Nilsenová (marie.postma@uvt.nl)
Department of Cognitive Science and Artificial Intelligence, Tilburg University,
Warandelaan 2, 5037 AB Tilburg, The Netherlands

Abstract

Producing and processing speech involves complex feedback loops of sensory and motor signals. Vocal sounds are partially processed as a movement affordance, allowing us to learn speaking patterns through imitation, which can be beneficial for language learning. In this study, we examine this process as a type of social embodiment illusion — the blurring of boundaries between self and other. Participants performed an altered version of a theatrical game called the 'one-voice expert', where they improvised speech in same-gender dyads. Unlike previous studies, we looked separately at the effects of simultaneousness ('speaking at the same time') and synchronicity ('saying the same thing'). These two variables were found to influence vocal characteristics and self-voice recognition in a distinct way, with synchronicity leading to stronger pitch adaptation and simultaneousness to suppression of phonetic convergence. We conclude that linking embodiment processes to joint speech in real world social interactions could be a promising new conceptual framework, with possible applications for language learning.

Keywords: voice; speech; social bonding; multisensory integration; phonetic convergence; embodiment; envoicement;

Introduction

The 'One-Voice Expert' (Keith, 1979) is a performative game originating from improvisational theater. In this game, two actors pretend to be one person, an expert on a certain made-up topic. They improvise answers to interview questions, speaking at the same time as if they have 'one voice'. To do this effectively, the speakers need to focus their attention fully on the other speaker and the task itself in order to quickly adapt the content of their utterances to the other, leading to a state of shared intentionality (Reddish, Fischer, & Bulbulia, 2013). A notable aspect of the game is that it involves joint speech that can be described as both *synchronous* (saying the same thing) and *simultaneous* (speaking at the same time) (Cummins, 2002). This kind of vocal activity can only be found in limited contexts in daily life, such as in choir singing or chanting.[1] These typically lead to an increase in social bonding (Mogan, Fischer, & Bulbulia, 2017) and a change from self- to we-agency (Salmela & Nagatsu, 2017).

Joint Speech as Joint Movement

Producing and monitoring speech is a process of multisensory integration. Afferent motor commands, as well as motor and sensory feedback, are an integral part of the experience of speaking (Postma, 2000). Additionally, there are functional links between the motor cortex and language systems in the brain, where words on a semantic level are partially processed according to their movement affordance (Pulvermüller, Hauk, Nikulin, & Ilmoniemi, 2005). Rhythmic sounds produced by the body, such as clapping or singing, are processed as not only an auditory signal but also as an intentional sequence of motor acts (Overy & Molnar-Szakacs, 2009). Even non-melodic speech contains rhythm and organized sound sequences. This allows listeners to mimic, and synchronize with, the vocal utterances of others (Cummins, 2009).

'Shadowing' speech in this manner provides another layer of processing that allows the listener to learn speech more quickly (Kadota, 2019). Listeners who subconsciously align their speech rhythm with another voice have an advantage at new-word learning tasks, and show structural differences in speech production and perception areas of the brain (Assaneo et al., 2019). Moreover, talented language learners have been shown to adapt their speech characteristics more strongly to others (Lewandowski & Jilka, 2019). Speaking with others can be done in several ways, and understanding those better allows us to determine the most effective learning strategies.

Synchronous vs. Simultaneous

Speech synchronization is a complex, adaptive process between speakers. In joint speech research, dialogue (which is both asynchronous and alternating) is often compared directly with singing (Kreutz, 2014), close shadowing, (Pardo et al., 2018), or non-improvised, joint synchronous speech (Cummins, 2002) (all both synchronous and simultaneous).

In non-improvised, joint synchronous speech, two participants read a text together, leading to mutual accommodation of speech patterns. Speakers use a variety of auditory information to optimize synchronization (Cummins, 2003, 2009). Similarly, both in close speech shadowing (quick synchronization with a recorded voice (Chistovich, Fant, de Serpa-Leitao, & Tjernlund, 1966; Marslen-Wilson, 1985)) and in 'normal' conversational speaking, lasting convergence or adaptation effects were found that persisted after the speaking tasks. In dialogues, this seems to be governed mainly by mimicry effects on pitch (Gijssels, Casasanto, Jasmin, Hagoort, & Casasanto, 2016). However, the vocal characteristics where these effects occur differ, and findings are inconsistent over different studies (Pardo et al., 2018).

Synchronized movement, such as marching or dancing, is

[1] Conversely, the most common form of joint speech - dialogue - is primarily asynchronous and alternating, with the exception of brief moments where speakers overlap or mimic each other.

a very common social practice (Wiltermuth & Heath, 2009) that influences social bonding — for example, the pro-social behavior of young children can be improved by dancing together (Kirschner & Tomasello, 2010). Vocal behavior such as joint speaking, chanting, or singing falls under this same category and can likewise lead to quick social bonding (Pearce, Launay, & Dunbar, 2015; Pearce et al., 2016). These effects of collective motor behavior might be closely related to the underlying mechanisms of embodiment illusions and the blurring of self-other boundaries (Tarr, Slater, & Cohen, 2018; Rombout, Atzmueller, & Postma-Nilsenová, 2018).

Embodiment Illusions

The One-Voice Expert game could be considered a new, social type of embodiment illusion. Embodiment illusions, such as the rubber hand illusion (Botvinick & Cohen, 1998), are commonly used to study how the brain demarcates agency and body-ownership, necessary for a number of fundamental processes including self-awareness and social interaction.

These illusions have been shown to elicit various effects. The embodied form can influence our beliefs and subsequent behaviors, even after short periods of embodiment (Banakou, Groten, & Slater, 2013; Maister, Sebanz, Knoblich, & Tsakiris, 2013). Within enfacement studies, synchronous stimulation has been shown to affect self-other recognition, inclusion of other in the self, judgment of resemblance and attractiveness, and affective state (Tsakiris, 2008; Sforza, Bufalari, Haggard, & Aglioti, 2010; Mazzurega, Pavani, Paladino, & Schubert, 2011; Maister, Banissy, & Tsakiris, 2013).

Embodiment illusions are generally assumed to arise through multisensory integration.[2] The predictive coding theory of embodiment states that the brain aims to minimize prediction errors that arise from integrating several sensory inputs, weighting them to arrive at a flexible model of what the current 'self' looks like (Kilteni, Maselli, Kording, & Slater, 2015). Visual feedback is generally weighted more heavily than other sensory input, and could even be essential for identification of the self (Tsakiris, 2017). However, it is not unthinkable that a combination of other sensory inputs could be just as strong. We would also expect individual differences in how sensory channels are weighted (Suzuki, Garfinkel, Critchley, & Seth, 2013; Tajadura-Jiménez & Tsakiris, 2014).

Speech and the Self-other Boundary

Distinguishing your own voice from others might be governed by the same mechanisms that play a role in body- and face-recognition (Graux, Gomot, Roux, Bonnet-Brilhault, & Bruneau, 2014). The voice is sometimes referred to as the 'auditory face', carrying information about identity and affect (Belin, Fecteau, & Bédard, 2004). Interestingly, when a person is speaking in synchrony with someone else, their vocal utterances are treated by the brain as if they derive from the other (Jasmin et al., 2016). This is surprising, as self-voice recognition has been shown to be quite robust (Xu, Homae, Hashimoto, & Hagiwara, 2013).

In one of the few voice-based embodiment illusions that have been studied so far, the addition of a speaking illusion to a virtual full-body illusion showed that people can attribute speech to themselves when they seem to inhabit the body that produces the speech (Banakou & Slater, 2014; Tajadura-Jiménez, Banakou, Bianchi-Berthouze, & Slater, 2017). The speaking illusion was created by combining auditory feedback with synchronous vibrotactile stimulation on the throat. This 'envoicement' illusion can have an immediate effect on participants' vocal output afterwards, similar to the effects of joint speech, although results have been mixed (Banakou & Slater, 2017). Additionally, two studies have looked at this 'rubber voice' illusion in isolation, and found that people interpreted a stranger's voice as their own if they heard it while speaking the same word themselves (Zheng, MacDonald, Munhall, & Johnsrude, 2011), but that auditory and vibrotactile feedback on their own might not be sufficient to elicit the illusion (Rombout & Postma-Nilsenova, 2019). These studies strongly suggest that envoicement effects might play a role in real-life social interactions, especially in joint speech situations.

In this study, we use the conceptual framework of embodiment to create a new type of social envoicement illusion. We believe this context could offer new perspectives on joint speech and group cognition, as well as — ultimately — on social mechanisms for language learning. We explicitly separate simultaneousness and synchronicity of speech, as they might cause different forms of vocal adaptation. Perhaps this could contribute to an explanation as to why results on phonetic convergence differ over different studies. If embodiment effects do indeed play a role in synchronous motor behavior and social interaction, blurring of the boundaries between self and other would be expected to occur — more so during speech that is both synchronous and simultaneous. In contrast, simultaneous yet asynchronous speech may disrupt any envoicement effects due to the 'erroneous' feedback this would cause (Alimardani et al., 2013; Rombout & Postma-Nilsenova, 2019).

Methods

Sample Sizes

Sixty-six participants (40 female, 26 male, average age 23 years (sd = 3.2)) formed 33 gender-matched dyads, separated in N = 16 for the synchronous/simultaneous condition (C1), N = 18 for asynchronous/simultaneous (C2), N = 16 for synchronous/alternating (C3), and N = 16 for asynchronous/alternating (C4). For 4 participants, the sound recordings were compromised, leaving N = 62 for the reaction time and voice characteristics measures (resulting in N = 16 for C1, N = 16 for C2, N = 14 for C3, and N = 16 for C4).

The subjects were all university students, gathered from the human subjects pool of Tilburg University in the Netherlands,

[2] However, visual feedback with only efferent motor commands seems to be sufficient (Alimardani, Nishio, & Ishiguro, 2013).

and native Dutch speakers, with no self-reported speech or hearing issues. They were rewarded with study credits. Participation was voluntary and informed consent was obtained from all subjects. The study was approved by the Tilburg Research Ethics and Data Management Committee.

Conditions and the interview-paradigm

The interviews were based on the 'One-Voice' Expert theatrical game. Two interviews were conducted per dyad, with questions about made-up areas of expertise — 'catching spears' and 'eating habits of odd ducks'. The questions were open-ended, such as: "What is the most peculiar eating habit of odd ducks?" and "What equipment do you need for catching spears?". The interviews were conducted in Dutch. All dyads were advised to speak slowly, repeat the question as part of their answer, and speak in full sentences.

The study used a 2x2 between subjects design. Conditions were assigned randomly and counterbalanced over dyads. Condition 1 can be described as synchronous / simultaneous. Participants were instructed to answer the interview-questions at the same time, as if they were a single speaker, without anyone taking the lead. In condition 2, asynchronous / simultaneous, participants answered the questions at the same time, but gave different answers, while keeping their answers approximately the same length. In condition 3, synchronous / alternating, participants alternated their answers (with the experimenter indicating who had to answer first); the second speaker was instructed to answer after the first speaker and repeat their answer exactly. Finally in condition 4, asynchronous / alternating, subjects alternated their answers, and gave different answers of approximately the same length (see Figure 1).

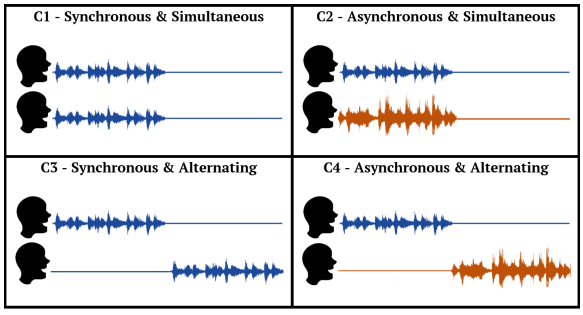

Figure 1: The four interview conditions. In the synchronous conditions (C1 and C3) participants say the same thing. In the simultaneous conditions (C1 and C2), they speak at the same time.

Measurements

Vocal Characteristics A list of 30 syllables was recorded by the participants both before and after the experiment to allow for comparison of vocal characteristics. The syllables were made up of a consonant and two vowels ('Foo', 'Kaa'), and consisted of common Dutch syllables that have no particular meaning on their own. Changes in vocal characteristics after the interviews were then determined by comparing these recordings.

Voice-recognition task A timed forced-choice task measured the strength of the vocal self/other boundary (Rombout & Postma-Nilsenova, 2019). 20 syllables were presented, and for each one the participant had to determine as fast as possible whether they were hearing their own voice, or the voice of someone else. Three tests were conducted for each participant - one before the interviews, one after the first interview, and one after the second interview. Each task consisted of 10 own-voice and 10 other-voice samples, presented in a random order. The 30 syllables recorded to compare vocal characteristics were also used for these stimuli, normalized to -5 decibel and with background noise removed.

AffectButton Possible changes in the participants' subjective valence, arousal and dominance after the social interaction were measured using the AffectButton (Broekens & Brinkman, 2013), a tool that translates these three dimensions to a more intuitive visual representation of an emoticon. Participants moved their mouse to change the facial expression of the emoticon until it represented the way they felt. We then calculated the delta between before and after the interviews.

Embodiment/Envoicement To measure envoicement, we adapted the original embodiment questionnaire (Botvinick & Cohen, 1998) in such a way that all statements referred to the voice. After the interviews, participants indicated their agreement with the statements on a 7-point Likert scale.

Affiliation Questionnaires Closeness was measured after the interviews using the IOS (inclusion of other in the self) scale (Woosnam, 2010). Additionally, the participants were asked to indicate on a 7-point Likert scale how attractive they thought the other participants' voice was, how well the collaboration went, and whether they considered the other participants' voice similar to their own in terms of pitch, timbre and rhythm. These were all based on similar questionnaires used in enfacement research (Sforza et al., 2010).

Procedure

Two soundproof booths were used for concurrent testing with two desktop computers and two Sennheiser headsets with microphones. The experiment was run in OpenSesame, with the exception of the interview portions, which were run in the voice-chat program TeamSpeak.

First, participants answered demographic questions and the AffectButton. Then, the participants recorded the 30 syllables and performed the first voice-recognition test. After this they were instructed to switch to the voice-chat, where the first interview took place. The interview questions appeared written on the screen and the participants answered them out loud, while hearing each-other speak through their headphones. Each dyad answered the questions according to the condition they were randomly assigned to. The interviews lasted from 3 to 5 minutes. After the first interview, participants did another voice-recognition test, then the second interview and a last voice-recognition test. They then filled out the IOS scale,

the second AffectButton, the envoicement questionnaire, general and manipulation-check questions, and lastly recorded the same 30 syllables again.

Table 1: Descriptive statistics and model-comparisons

	Mean	SD	$\chi^2(1)$	p-value
f0				
Sync	0.92	22.13		
Async	3.51	27.7	3.86	0.049
f1				
Sim	12.39	209.43		
Alt	30.67	197.58	6.13	0.013
f1 (interaction)				
Sync/Sim	28.6	204.64		
Async/Sim	3.82	213.09		
Sync/Alt	20.88	167.86		
Async/Alt	39.24	220.14	8.90	0.031
f2				
Sim	23.37	330.66		
Alt	42.87	327.09	6.05	0.014
HNR				
Sim	0.18	1.64		
Alt	0.1	1.77	3.93	0.047

Results

Vocal Characteristics

All sound-samples of the syllables recordings (60 per participant) were analyzed with the Soundgen library in R (R Core Team, 2013; Anikin, 2018), using a PitchFloor of 50 and a PitchCeiling of 500. The fundamental frequency of these voice-samples was extracted to calculate the possible f0-shift in reaction to the other participant. Additionally, the convergence on f1, f2 and the harmonics-to-noise ratio (HNR) was also calculated to account for secondary voice characteristics. This individual adaptation was calculated per syllable by taking the delta between the before measurement from the participant and that of their dyad counterpart, and subtracting the delta between the after measurement of the participant and the before measurement of their dyad counterpart (Postma-Nilsenová, Brunninkhuis, & Postma, 2013). Thus, only identical syllables were directly compared.

We used R with the lme4 library (R Core Team, 2013; Bates, Maechler, Bolker, & Walker, 2012) to perform a linear mixed effects analysis of the relationship between pitch, f1, f2 or HNR, and a synchronous interview, a simultaneous interview, or the interaction of both. We added gender as a fixed effect and included intercepts for subjects, syllables and dyads as random effects.[3] Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. P-values were obtained by likelihood ratio tests of the full model against the basic model. We only report the significant differences, see table 1. Detailed output of the mixed models can be found in table 2.

The synchronous conditions, i.e., producing an utterance with the same content, only had a significant effect on f0 adaptation, a higher mean indicating more adaptation. On the other hand, the simultaneous conditions influenced the secondary vocal features. These results show that after the conditions where participants spoke at the same time, they adapted their vocal characteristics less to each other than after the conditions where they took turns. Additionally, there was a significant interaction effect on f1, with least adaptation after the sync/sim (C1) condition and most after the async/alt (C4) condition.

Voice Recognition task

We did not find systematic effects of the experimental conditions on the overall voice recognition score. Participants became faster over time (F(1, 57) = 35.103, $p < .001$, $\eta p2 = .381$). There was no significant interaction effect between time and condition. However, during the second reaction-time task (between the two interviews), participants made significantly more mistakes in the simultaneous conditions (F(1,60) = 5,227, p = .026), and were significantly quicker in the synchronous conditions (F(1,59) = 4,692, p = .034).

Table 2: Summary of Mixed Effects Models

	Estimate	SE	t-value
f0			
Intercept	-5.273	2.170	-2.430
Gender (female)	2.818	2.313	1.218
Sync/Async (sync)	4.502	2.255	1.997
f1			
Intercept	30.5740	15.6829	1.950
Gender (female)	0.1756	16.9762	0.010
Sim/Alt (sim)	-43.0842	16.5463	-2.604
f1 (interaction)			
Intercept	39.945	18.137	2.202
Gender (female)	-1.418	16.325	-0.087
Sync/Async (sync)	-18.258	22.527	-0.811
Sim/Alt (sim)	-35.061	22.113	-1.586
Sync:Sim	-14.337	31.444	-0.456
f2			
Intercept	56.69	23.02	2.463
Gender (female)	-25.92	24.70	-1.049
Sim/Alt (sim)	-62.24	24.08	-2.585
HNR			
Intercept	-0.07767	0.14652	-0.530
Gender (female)	0.33026	0.16744	1.972
Sim/Alt (sim)	-0.32872	0.16320	-2.014

[3] Basic model (example for pitch (f0)) created as follows: $model0 = lmer(pitch\_adaptation \sim gender + (1|pp) + (1|syll) + (1|dyad), data = datafull, REML = FALSE)$. Full model created by adding *sync*, *sim* or $sync * sim$.

Other Measurements

There were no significant differences between conditions in the delta of the valence, arousal and dominance scores, in the IOS score, on the embodiment questionnaire, or on the additional questionnaires. There was an effect on the subjective experience of the collaboration ($F(3,60) = 11.691$, $p < .001$). Participants rated the collaboration significantly higher in C3 (sync/alt) than in C1 (sync/sim) (Tukey post-hoc test, C3: 6.6 ± 1.1, C1: 5.2 ± 1.4, $p = .023$). Additionally, participants in alternating conditions (3 and 4) rated the collaboration significantly higher than in C2 (async/sim) (Tukey post-hoc test, C3: 6.6 ± 1.1, C2: 3.9 ± 1.8, $p < .001$) (Tukey post-hoc test, C4: 5.9 ± 1.1, C2: 3.9 ± 1.8, $p = .001$).

Discussion

We proposed a new social envoicement illusion, aiming to study joint speaking in the context of self-other boundaries. We specifically separated synchronous and simultaneous joint speech so these characteristics could be studied in isolation, especially in regards to their effects on phonetic convergence (Pardo et al., 2018). Additionally we expected similar blurring of self and other as, for example, in the enfacement illusion (Tsakiris, 2008), with strongest effects after joint speech that was both synchronous and simultaneous, and little effect after simultaneous yet asynchronous speech.

The analysis of voice characteristics revealed that synchronicity and simultaneousness might indeed affect vocal adaptation differently. In the conditions where the participants were instructed to say the same thing, they adapted their pitch significantly more towards the other person. This could indicate a blurring of body-boundaries or a stronger speech mimicking impulse, which could be associated with stronger language learning outcomes (Assaneo et al., 2019). This suggests possibilities for practical applications of embodiment/envoicement paradigms.

Speaking at the same time, on the other hand, had a significant effect on several secondary voice characteristics. When the participants were taking turns talking, their voices became more similar, which is likely due to a mimicry effect as is expected in normal conversation (Postma-Nilsenová et al., 2013; Gijssels et al., 2016). But when talking at the same time, this effect seems to be suppressed and the distance between voices stays the same or even diverges. A possible explanation could be that this helps to distinguish the own voice from the other, avoiding confusion and slurring of speech (Marslen-Wilson, 1985). These results confirm the complex nature of how the voice is influenced by joint speech (Pardo et al., 2018), reflected in the novel separation of synchronous and simultaneous speaking. It seems very likely that other, similar joint speech features may play a role as well.

The results for embodiment effects, or blurring between self and other, were more mixed. Performance on the voice recognition task improved over time, which is most likely a training effect. After the first interview, participants in the synchronous conditions specifically were significantly faster at the recognition task, an effect that disappeared after the second interview. One explanation could be that the utterances with the same content made the differences between the two voices stand out more, as participants could compare them quite directly with each-other. In the simultaneous conditions on the other hand, participants made significantly more mistakes on the recognition task after the first interview, an effect that again disappeared after the second. It is possible that the act of speaking at the same time made it more difficult for participants to distinguish their voices from each other (which would tie in with the effect this had on secondary voice characteristics). In both cases it is however unclear why these effects did not persist.

The remaining questionnaire results indicate that on the subjective level, there was little difference in how the participants experienced the different conditions (except on collaboration satisfaction, which is unlikely to be an embodiment effect). The enfacement illusion quite reliably influences subjective experience (Tsakiris, 2008), so this is an indication that an envoicement illusion without any visual components might not be strong enough to do so. This is not entirely unexpected, as vision is generally recognized as a much stronger contributor to embodiment illusions (Tsakiris, 2017; Rombout & Postma-Nilsenova, 2019). The effects on voice characteristics and self-voice recognition show that some blurring of self/other boundaries may have occurred on a more subconscious level, but not always in the direction that we would have expected beforehand. To explore this further, future research could try and strengthen any embodiment effects through an added visual illusion (for example a virtual reality body swap).

Conclusion

We explored a new experimental paradigm designed to elicit a social 'envoicement' illusion and study joint speech in the context of embodiment and self-other boundaries. Our results show that the synchronous and simultaneous qualities of joint speech influence vocal adaptation differently. Synchronous speech significantly strengthens f0 adaptation, but seems to also improve self-voice recognition, whereas simultaneous speech causes the adaptation of secondary voice characteristics to be suppressed, and self-voice recognition to be more difficult.

These results shows the importance of studying these joint speech features separately, and points to the possibility of other features that might influence phonetic convergence. Our results on the possible effects of embodiment processes in joint speech are only exploratory, but warrant further research. Uncovering the role of envoicement in social interaction and joint speech could not only lead to a better understanding of both, but ultimately to applications where one is strengthened by the other — for example by using virtual embodiment and voice illusions to improve language learning through increased vocal adaptation.

References

Alimardani, M., Nishio, S., & Ishiguro, H. (2013). Humanlike robot hands controlled by brain activity arouse illusion of ownership in operators. *Scientific reports*, *3*, 2396.

Anikin, A. (2018). Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods*, 1–15.

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., & Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature neuroscience*, *22*(4), 627–632.

Banakou, D., Groten, R., & Slater, M. (2013). Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *Proceedings of the National Academy of Sciences*, *110*(31), 12846–12851.

Banakou, D., & Slater, M. (2014). Body ownership causes illusory self-attribution of speaking and influences subsequent real speaking. *Proceedings of the National Academy of Sciences*, *111*(49), 17678–17683.

Banakou, D., & Slater, M. (2017). Embodiment in a virtual body that speaks produces agency over the speaking but does not necessarily influence subsequent real speaking. *Scientific reports*, *7*(1), 1–10.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2012). lme4: Linear mixed-effects models using eigen and s4 r package version 10-4 available: http. *CRANR-projectorg/package= lme4*.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, *8*(3), 129–135.

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, *391*(6669), 756.

Broekens, J., & Brinkman, W.-P. (2013). Affectbutton: A method for reliable and valid affective self-report. *Int. Journal of Human-Computer Studies*, *71*(6), 641–667.

Chistovich, L., Fant, G., de Serpa-Leitao, A., & Tjernlund, P. (1966). Mimicking of synthetic vowels. *Quarterly Progress and Status Report, Speech Transmission Lab, Royal Institute of Technology, Stockholm*, *2*, 1–18.

Cummins, F. (2002). On synchronous speech. *Acoustics Research Letters Online*, *3*(1), 7–11.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics*, *31*(2), 139–148.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, *37*(1), 16–28.

Gijssels, T., Casasanto, L. S., Jasmin, K., Hagoort, P., & Casasanto, D. (2016). Speech accommodation without priming: The case of pitch. *Discourse Processes*, *53*(4), 233–251.

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., & Bruneau, N. (2014). Is my voice just a familiar voice? an electrophysiological study. *Social cognitive and affective neuroscience*, *10*(1), 101–105.

Jasmin, K. M., McGettigan, C., Agnew, Z. K., Lavan, N., Josephs, O., Cummins, F., & Scott, S. K. (2016). Cohesion and joint speech: Right hemisphere contributions to synchronized vocal production. *Journal of Neuroscience*, *36*(17), 4669–4680.

Kadota, S. (2019). *Shadowing as a practice in second language acquisition*. Routledge.

Keith, J. (1979). Impro: improvisation and the theatre. *Londen: Faber and Faber Ltd*.

Kilteni, K., Maselli, A., Kording, K. P., & Slater, M. (2015). Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. *Frontiers in human neuroscience*, *9*.

Kirschner, S., & Tomasello, M. (2010). Joint music making promotes prosocial behavior in 4-year-old children. *Evolution and Human Behavior*, *31*(5), 354–364.

Kreutz, G. (2014). Does singing facilitate social bonding. *Music Med*, *6*(2), 51–60.

Lewandowski, N., & Jilka, M. (2019). Phonetic convergence, language talent, personality & attention. *Frontiers in Communication*, *4*, 18.

Maister, L., Banissy, M. J., & Tsakiris, M. (2013). Mirror-touch synaesthesia changes representations of self-identity. *Neuropsychologia*, *51*(5), 802–808.

Maister, L., Sebanz, N., Knoblich, G., & Tsakiris, M. (2013). Experiencing ownership over a dark-skinned body reduces implicit racial bias. *Cognition*, *128*(2), 170–178.

Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. *Speech communication*, *4*(1-3), 55–73.

Mazzurega, M., Pavani, F., Paladino, M. P., & Schubert, T. W. (2011). Self-other bodily merging in the context of synchronous but arbitrary-related multisensory inputs. *Experimental brain research*, *213*(2-3), 213–221.

Mogan, R., Fischer, R., & Bulbulia, J. A. (2017). To be in synchrony or not? a meta-analysis of synchrony's effects on behavior, perception, cognition and affect. *Journal of Experimental Social Psychology*, *72*, 13–20.

Overy, K., & Molnar-Szakacs, I. (2009). Being together in time: Musical experience and the mirror neuron system. *Music Perception*, *26*(5), 489–504.

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, *69*, 1–11.

Pearce, E., Launay, J., & Dunbar, R. I. (2015). The ice-breaker effect: singing mediates fast social bonding. *Open Science*, *2*(10), 150221.

Pearce, E., Launay, J., van Duijn, M., Rotkirch, A., David-Barrett, T., & Dunbar, R. I. (2016). Singing together or apart: The effect of competitive and cooperative singing on social bonding within and between sub-groups of a university fraternity. *Psychology of music*, *44*(6), 1255–1273.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*,

77(2), 97–132.

Postma-Nilsenová, M., Brunninkhuis, N., & Postma, E. (2013). Eye gaze affects vocal intonation mimicry. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, *21*(3), 793–797.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Reddish, P., Fischer, R., & Bulbulia, J. (2013). Lets dance together: synchrony, shared intentionality and cooperation. *PloS one*, *8*(8), e71182.

Rombout, L. E., Atzmueller, M., & Postma-Nilsenová, M. (2018). Towards estimating collective motor behavior: Aware of self vs. aware of the other. *Proceedings of the Workshop on Affective Computing and Context Awareness in Ambient Intelligence*.

Rombout, L. E., & Postma-Nilsenova, M. (2019). Exploring a voice illusion. In *2019 8th international conference on affective computing and intelligent interaction (acii)* (pp. 711–717).

Salmela, M., & Nagatsu, M. (2017). How does it really feel to act together? shared emotions and the phenomenology of we-agency. *Phenomenology and the Cognitive Sciences*, *16*(3), 449–470.

Sforza, A., Bufalari, I., Haggard, P., & Aglioti, S. M. (2010). My face in yours: Visuo-tactile facial stimulation influences sense of identity. *Social neuroscience*, *5*(2), 148–162.

Suzuki, K., Garfinkel, S. N., Critchley, H. D., & Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia*, *51*(13), 2909–2917.

Tajadura-Jiménez, A., Banakou, D., Bianchi-Berthouze, N., & Slater, M. (2017). Embodiment in a child-like talking virtual body influences object size perception, self-identification, and subsequent real speaking. *Scientific Reports*, *7*(1), 1–12.

Tajadura-Jiménez, A., & Tsakiris, M. (2014). Balancing the inner and the outer self: Interoceptive sensitivity modulates self–other boundaries. *Journal of Experimental Psychology: General*, *143*(2), 736.

Tarr, B., Slater, M., & Cohen, E. (2018). Synchrony and social connection in immersive virtual reality. *Scientific reports*, *8*(1), 3693.

Tsakiris, M. (2008). Looking for myself: current multisensory input alters self-face recognition. *PloS one*, *3*(12), e4040.

Tsakiris, M. (2017). The multisensory basis of the self: from body to identity to others. *The Quarterly Journal of Experimental Psychology*, *70*(4), 597–609.

Wiltermuth, S. S., & Heath, C. (2009). Synchrony and cooperation. *Psychological science*, *20*(1), 1–5.

Woosnam, K. M. (2010). The inclusion of other in the self (ios) scale. *Annals of Tourism Research*, *37*(3), 857–860.

Xu, M., Homae, F., Hashimoto, R.-i., & Hagiwara, H. (2013). Acoustic cues for the recognition of self-voice and other-voice. *Frontiers in psychology*, *4*.

Zheng, Z. Z., MacDonald, E. N., Munhall, K. G., & Johnsrude, I. S. (2011). Perceiving a stranger's voice as being one's own: A rubber voice illusion? *PloS one*, *6*(4), e18655.