Manuscript title: Speech motor adaptation during synchronous and metronome-timed speech

Abigail R. Bradshaw[1,2]*

Daniel R. Lametti[3]

Douglas M. Shiller[4]

Kyle Jasmin[5]

Ruiling Huang[1]

Carolyn McGettigan[1]

[1]Department of Speech, Hearing and Phonetic Sciences, University College London, London, UK

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

[3]Department of Psychology, Acadia University, Wolfville, Nova Scotia, Canada

[4] School of Speech-Language Pathology and Audiology, l'Université de Montréal, Montreal, Canada

[5] Department of Psychology, Royal Holloway, University of London, London, UK

*Corresponding author information:

Dr Abigail Bradshaw, MRC Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Road, Cambridge, CB2 7EF, UK (email: abbie.bradshaw@mrc-cbu.cam.ac.uk).

Word count: 7919

Author note: Data and analysis scripts for the experiments reported on in this manuscript are openly available on the Open Science Framework and can be accessed at: https://osf.io/gz87s/. The design, hypotheses and analyses for Experiment 1 were pre-registered (https://osf.io/q86cj). The ideas and data appearing in this manuscript were presented at the 8th International Conference on Speech Motor Control and the 14th Annual Meeting of the Society for the Neurobiology of Language.

Abstract

Sensorimotor integration during speech has been investigated by altering the sound of a speaker's voice in real-time; in response, the speaker learns to change their production of speech sounds in order to compensate (adaptation). This line of research has however been predominantly limited to very simple speaking contexts, typically involving (1) repetitive production of single words and (2) production of speech whilst alone, without the usual exposure to other voices. This study investigated adaptation to a real-time perturbation of the first and second formants during production of sentences either in synchrony with a pre-recorded voice (synchronous speech group) or alone (solo speech group). Experiment 1 (*n = 30*) found no significant difference in the average magnitude of compensatory formant changes between the groups; however, synchronous speech resulted in increased between-individual variability in such formant changes. Participants also showed acoustic-phonetic convergence to the voice they were synchronising with prior to introduction of the feedback alteration. Furthermore, the extent to which the changes required for convergence agreed with those required for adaptation was positively correlated with the magnitude of subsequent adaptation. Experiment 2 tested an additional group with a metronome-timed speech task (*n = 15*), and found a similar

pattern of increased between-participant variability in formant changes. These findings demonstrate that speech motor adaptation can be measured robustly at the group level during performance of more complex speaking tasks; however, further work is needed to resolve whether self-voice adaptation and other-voice convergence reflect additive or interactive effects during sensorimotor control of speech.

Keywords: sensorimotor learning; speech adaptation; synchronous speech

Public significance statement

- The current paper extends previous research on the role of speech auditory feedback (the sound of our voice as we are speaking) in speech motor control by investigating it within more complex speaking contexts.
- We found that compared to natural sentence reading, both speaking in synchrony with (1) another person and (2) a metronome-beat resulted in more variable changes in speech production when exposed to experimentally induced ‘errors’ in real-time speech auditory feedback.
- These results suggest the need to consider how other processes relevant to the current speaking context (e.g. perception of external voices) may interact or be combined with speech motor learning in response to auditory errors.

Previous research on speech motor control has highlighted the importance of monitoring the sound of our own voice as we are speaking (termed ‘speech auditory feedback’) for speech production. This has been demonstrated in altered auditory feedback experiments, in which speakers exposed to real-time alterations of their voice as they are speaking are found to show unconscious compensatory changes in the way they produce speech sounds (Burnett et al., 1998; Houde & Jordan, 1998). For example, by altering the resonant frequencies (known as formants) of a speech signal, subtle changes can be made to the vowel sound e.g., alteration of the first and second formants can result in a produced “head” sounding more like “had”. Consistent exposure to such a formant alteration results in compensatory changes in produced formants e.g. so that something closer to “hid” is produced. This ‘speech motor adaptation’ is explained by models of speech motor control proposing that the brain uses forward models to predict the sensory consequences of a speech gesture; these predictions allow for potential errors to be detected, which can drive compensatory adjustments to the forward models via sensorimotor learning (Guenther et al., 2006; Houde & Nagarajan, 2011; Parrell, Lammert, et al., 2019). The persistence of these compensatory changes once the feedback alteration has been removed (termed the after-effects of adaptation) have been used as key evidence that such sensorimotor learning has occurred (Purcell & Munhall, 2006).

However, these models currently do not include in their scope the influence of other voices on speech production. This absence likely reflects the fact that much previous research with the altered auditory feedback paradigm has involved producing speech alone, typically in the context of single word repetition. This is a poor representation of real-world speech, which is characterised by the production of fluid, variable sentences, usually in the context of a social interaction involving exposure to at least

one other voice. It is already well documented that exposure to other voices can affect our own speech production. This is observed in the phenomenon of vocal convergence; a speaker's tendency to acquire (i.e. converge to) the acoustic-phonetic properties of a voice they are interacting with (Pardo et al., 2017). Although typically viewed in terms of a social attunement process (Giles et al., 1991), others have argued that vocal convergence can be viewed as a more basic level sensorimotor learning process (Sato et al., 2013). That is, exposure to another voice might, like altered auditory feedback, trigger an updating of stored internal sensorimotor models that then drive changes in speech production. Viewed from this perspective, a question then arises as to how adaptation and convergence might interact with one another.

So far, only a small number of studies has looked at how experience of other voices affects adaptation, although none of these has specifically looked at vocal convergence. A study by Lametti et al., (2014) demonstrated that explicit perceptual training that aimed to shift perception of a phoneme boundary when listening to another voice affected the magnitude of subsequent speech motor adaptation during production of those same vowels. This perceptual training effect has also been reported in a sample of children (Shiller & Rochon, 2014). This was taken even further in a study by Bourguignon et al., (2016), who demonstrated a similar type of effect caused by mere incidental exposure to another voice. Specifically, they found that manipulating the first formant of a voice that spoke a cueing phrase to be either high or low had systematic effects on the magnitude of a participant's speech adaptation response to an F1 perturbation on a subsequent word production. A study by Sato & Shiller (2018) failed to find a significant difference in adaptation responses between an auditory-cued and a visual-cued condition; however, in this study the

auditory tokens used to cue speech productions were in fact recordings of each participant's own speech productions (without any systematic manipulations of their spectral properties). Altogether therefore, this preliminary evidence suggests that speech motor adaptation can be sensitive to the acoustic-phonetic properties of other voices in our speech environment.

In the current study, we aimed to investigate adaptation in the context of a speech task involving the production of variable sentences that preserves and enhances this aspect of everyday speech (that is, the experience of other voices), by combining an altered auditory feedback paradigm with a synchronous speech task. Synchronous speech refers to the act of speaking in unison with another speaker. While this task is clearly not representative of most forms of everyday speech, this behaviour is observed across a variety of real-world settings such as in places of worship and sports stadiums (Cummins, 2018). In addition, there are a number of aspects to this task which make it particularly well-suited to targeting questions concerning the role of other voices in speech sensorimotor processing. During synchronous speech, speech feedback from the self-voice and from another voice is received concurrently, forcing the parallel processing of both input streams at the same time. This provides an interesting context within which to explore the dynamics of these two sources of feedback on speech motor control. It has previously been demonstrated that speakers engaged in synchronous speech show convergent changes in the fundamental frequency (the acoustic correlate of pitch) of their speech productions towards that of the voice they are synchronising with (Bradshaw & McGettigan, 2021). Furthermore, there is neural evidence that synchronous speech may change the way in which the brain responds to self-produced speech feedback, resulting in a release from speech induced-suppression: the typical reduction in the auditory

response to a speech signal when it is actively produced versus passively listened to (Jasmin et al., 2016). This suggests that synchronous speech may have implications for sensorimotor processing of self-produced speech feedback.

By measuring adaptation during synchronous speech, we can thus investigate whether adaptation is affected by a speech task that involves both a change in speaking style and exposure to another voice. Speech adaptation has already been demonstrated to be robust when concurrently performing a visuo-motor task (Lametti et al., 2020), but it is unknown whether this would also hold during concurrent performance of a *speech* task that requires speakers to attend to and match the rhythm of another voice's speech. Aside from these additional task demands, we can also ask whether the exposure to the acoustic-phonetic properties of the other voice might disrupt adaptation, by engaging vocal convergence processes.

In this study, we used a modified version of the sentence-level adaptation task developed by Lametti et al., (2018), to compare adaptation responses to a formant perturbation between a group who produced sentences on their own (solo speech group) and a group who produced sentences in synchrony with another voice (synchronous speech group). We hypothesised that we would see a significant difference in the magnitude of adaptation to an auditory feedback perturbation between these groups, in the form of either reduced or enhanced adaptation in the synchronous speech group. Adaptation may be weakened because of the additional task demands or the presence of another voice; alternatively, it may be enhanced, perhaps by increasing reliance on feedback control over feedforward control. In Experiment 2, we additionally tested the effects of metronome-timed speech on adaptation. This provides a condition involving a similar speech task to synchronous

speech (i.e., synchronising the timing of one's speech with an external rhythm), but crucially in the absence of exposure to another voice.

Methods

This study received ethical approval from the local ethics officer at the Department of Speech, Hearing and Phonetic Sciences at University College London (approval no. SHaPS-2019-CM-030). All participants gave informed consent prior to taking part in the study.

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in this study. The design, hypothesis and analysis for Experiment 1 were pre-registered on the Open Science Framework (https://osf.io/q86cj). Data and analysis scripts for both experiments are openly available on the Open Science Framework and can be accessed at: https://osf.io/gz87s/. Data were analysed using R, version 3.6.1 (R Core Team, 2019), and the package lmerTest (Kuznetsova et al., 2017).

Experiment 1

Participants

Thirty participants (*mean age* = 22.53 years, *sd* = 4.15) took part in this experiment. We restricted our recruitment to females from the South-East of England so as to match with the gender and accent of the accompanist voice used for the synchronous speech task (see *Stimuli*). Demographic information on participants was collected via a self-report questionnaire. Participants were asked to tick one of four boxes (female, male, other or prefer not to say) in response to the prompt "Are

you: ". All participants recruited for this study ticked the box for female. Participants were asked to report their native language via a free response box. Finally, they were asked to tick yes or no in response to the question "Did you learn British-English as your native language within the South East of England (includes London, Berkshire, Buckinghamshire, Essex, Surrey, Hertfordshire, Kent, Hampshire, Sussex, Bedfordshire, Cambridgeshire, Oxfordshire)?". All participants reported being native speakers of British English from the South-East of England. No participants reported a history of hearing loss or a history of speech, language or reading disorders.

An equal number of participants took part in the solo speech and synchronous speech conditions (15 in each). This sample size was chosen based on previous work in which significant sensorimotor adaptation in response to feedback perturbations is seen at the group level in groups of 8-12 participants (Houde & Jordan, 1998; Lametti et al., 2018). We have chosen a sample size slightly higher than this per group, to increase sensitivity to potentially smaller effect sizes.

Procedure

Participants were asked to read sentences presented on a computer screen, either in a natural speaking manner (solo speech condition) or in synchrony with a pre-recorded voice (the accompanist) that they heard through headphones (synchronised speech condition), in a between-groups design. All participants completed 7 blocks of the same set of 50 sentences (see Figure 1A). For the synchronised speech group, the first block of sentences was read normally without the accompanist voice (as in the solo speech condition). Synchronised speech was then introduced from block 2 for this group. Both groups were given 10 practice trials

of solo speech at the start of the experiment, to familiarise them with the procedure and allow them to practice maintaining a speaking intensity within an acceptable range (see *Apparatus*). The synchronised speech group then underwent a further 10 practice trials of synchronised speech after the first baseline block.

Across both conditions, each trial began with presentation of a sentence onscreen and a three-click countdown through the headphones (with an interstimulus interval of 1 second between clicks). For the synchronised speech condition (blocks 2-7), this was followed by presentation of a pre-recorded voice speaking the sentence (the accompanist). Participants in the solo speech condition (and in the synchronised speech condition on block 1) were told to read the sentence after the countdown-clicks. Participants in the synchronised speech condition from block 2 onwards were told to use the three-click countdown to help them to start speaking at the same time as the accompanist, and to synchronise their speech-timing with them as closely as possible. Each trial lasted 8.5 seconds; each sentence remained onscreen for the full duration of the trial.

Stimuli

Sentence stimuli for the task were 50 sentences taken from the Harvard IEEE corpus of sentences (IEEE Subcommittee on Subjective Measurements, 1969). These sentences are designed to be phonetically balanced. Presentation of sentences was randomised within blocks. The accompanist voice audio stimuli used in the synchronised speech condition consisted of recordings of a female speaker of Standard Southern British English reading the same 50 sentences. These tokens were recorded using the internal microphone of a MacBook Air and the software

programme Audacity (Audacity Team, 2021). The audio recordings were matched for amplitude via RMS norming in Praat (Boersma & Weenink, 2021).

Apparatus

The set-up for the speech motor adaptation paradigm is illustrated in Figure 2. Participants produced sentences while wearing a headset (Beyerdynamic, DT 297 PV, Heilbronn, Germany) with circumaural headphones and a head-mounted microphone positioned to be approximately 10cm away from the participant's mouth. The signal from the microphone was routed to an audio interface (MOTU MicroBook IIc, Cambridge, MA) and then on to a mixer (Xenyx502, Behringer, Braintree, MA), where it was mixed with pink noise presented at 62dB (to mask the bone-conducted signal), and the audio file for that trial presented at 70dB e.g. the accompanist voice (see *Stimuli*). This mix was then fed back to the headphones. The level of the speech feedback through the headphones was calibrated by playing back a speech signal from a loudspeaker at a distance of 10cm from the microphone; the amplification level was then adjusted such that an input intensity of 77dB at the microphone (measured using a sound level meter Type 2231, Brüel & Kjær, Nærum, Denmark) also resulted in an output intensity of 77dB through the headphones. Calibration of all levels through the headphones was achieved using an artificial ear (Type 4153 with Type 4192 condenser microphone, Brüel & Kjær, Nærum, Denmark) connected to a PHOTON+ system using RT PRO software (Brüel & Kjær, Nærum, Denmark). During the task, participants were provided with visual feedback on their speaking level by an LED display (that received input from the audio interface) to help them maintain a target level between 75 and 80dB SPL as measured at the microphone (at a distance of 10cm from the participant's mouth). All participants were given 10 practice trials of solo speech at the start of the

experiment, in order to practise speaking at the correct intensity range, with feedback provided by the experimenter at this point.

Speech signal processing and manipulation was implemented by Audapter (Cai, 2015; Cai et al., 2008), a publicly available MATLAB-based application. Speech was recorded at a sampling rate of 48kHz (downsampled to 16kHz) with a buffer size of 96 samples. The feedback perturbation consisted of a joint F1-F2 perturbation implemented in mels; this is a perceptually normalized scale that takes into account the potentially large differences in F1 and F2 in Hz between vowels. Specifically, following Lametti, Smith, Watkins and Shiller (2018), produced F1 values were increased by 49.5 mels, and produced F2 values were decreased by 49.5 mels, resulting in a combined perturbation of 70 mels in F1/F2 space (see Figure 1B). Across both synchronised and solo speech groups, feedback was unaltered for the first two blocks of sentences; the feedback perturbation was then ramped up linearly over the first 25 trials of the third block (ramp phase) and then held constant over the last 25 trials of the third block and the next three blocks (hold phase), before being removed for the final block (after-effect phase) (see Figure 1C).

The total feedback loop latency of the audio set-up was measured using methods outlined in Kim et al., (2020). Using a value of 3 for the nDelay parameter within Audapter, the total feedback delay associated with both hardware and software latencies for perturbed speech feedback was measured at 26.68ms. This is almost identical to the latency value reported for an equivalent set-up by Kim et al., (2020), and is far below the delay levels that have been reported to disrupt speech adaptation (Max & Maffett, 2015; Shiller et al., 2020).

Acoustic Analysis

Speech was recorded at 16000 Hz in Matlab. Formant frequencies were analysed using a custom-made script in Praat (Boersma & Weenink, 2021). First, the vocalised/periodic portions of the acoustic signal were isolated using Praat's autocorrelation method (Boersma, 1993) (see Supplementary Material S1 for an example). Briefly, a signal (in this case, a brief window of recorded speech) is judged as periodic if its autocorrelation function reveals a maximum value above a certain threshold (e.g., 0.45). The unbiased autocorrelation algorithm used by Praat has been shown to be more precise and noise resistant than earlier autocorrelation methods or approaches based on comb filtering or cepstral analysis (Boersma, 1993). Linear Predictive Coding (LPC) analysis was then used to compute F1 and F2 values from these periodic segments. These were then averaged across each sentence within each production block.

Quantification of Adaptation

To look at participant responses to the formant perturbation, firstly a production change measure was calculated for F1 and F2 separately by calculating the change in produced formant frequencies from block 2 (baseline) to block 6 (the final adaptation block) for each sentence; the mean was then taken across the 50 sentences to give an average production change value for each participant (for each formant). Adaptation was then quantified using a measure which calculates the extent to which these changes in produced F1 and F2 directly countered the direction of the feedback perturbation in F1-F2 space (Lametti et al., 2018; Niziolek & Guenther, 2013). To do this, firstly the inverse of the vector representing the feedback shift in F1-F2 space was found; this vector represents perfect compensation to the feedback perturbation. The angular difference between this

inverse shift vector and a vector representing the participant's production change (relative to block 2) was then calculated; the cosine of this difference was then multiplied by the magnitude of production change. This measure of adaptation thus quantifies the degree to which the observed change in produced formants (i.e. the production change measure above) precisely opposed the feedback perturbation. This measure was calculated for each individual trial and then averaged within each block after the introduction of the feedback perturbation (blocks 3-7). Positive values on this measure indicate formant changes that opposed the direction of the perturbation; negative values reflect formant changes that followed the perturbation direction.

Hypotheses

Our main hypothesis for this experiment was that the magnitude of the adaptation response would be significantly different (either reduced or enhanced) in the synchronous speech group compared to the solo speech group. Thus, we predicted that while we would see a significant adaptation response at the group level in the solo speech group, adaptation may or may not be significant in the synchronous speech group.

Results

Pre-registered Analyses

Synchronisation success

To ensure that participants in the synchronous speech group were sufficiently performing the task of speech synchronisation, a Dynamic Time Warping algorithm was used to measure synchronisation success (Cummins, 2009; Ellis, 2003; Jasmin et al., 2016). Conceptually, this measures the level of synchrony between two

speaker's utterances by identifying the optimal warping of one onto the other. This measure was calculated for utterances from the second baseline block (block 2), to establish task engagement and success before the feedback perturbation was introduced.

Firstly, spectrograms of the participant's and accompanist's productions for a given trial were calculated. Next, a similarity matrix was created using the cosine distances of the spectrogram magnitudes; dynamic programming was then used to find the lowest-cost 'warp path' from one to the other (taking the accompanist's utterance as the 'referent' to be warped onto). This path was rotated to be expressed in the time domain, and rescaled into seconds. The unsigned area under the curve was then taken as a measure of asynchrony between the two utterances, with higher values thus representing poorer synchrony (Ellis, 2003). These values were averaged across trials within block 2 to give a mean asynchrony score for each participant. This was done for both solo speech and synchronous speech groups, in order to verify the expected difference in asynchrony scores according to which task was performed. An independent-samples t-test confirmed that asynchrony scores were significantly higher in the solo speech group (M = 0.401, SD = 0.101), compared to the synchronous speech group (M = 0.245, SD = 0.029): $t(16.29) = 5.76$, $p < .001$. This suggests that the between-subjects experimental manipulation had been successful, in that participants in the synchronous speech group were more synchronised to the accompanist voice than were the participants in the solo speech condition (who spoke alone and never heard the accompanist voice).

Adaptation responses

Individual adaptation vectors are illustrated in Figure 3A; individual vectors

representing the after-effects of adaptation are illustrated in Figure 3B. Adaptation responses quantified as the component of each participant's formant changes that directly opposed the perturbation are shown across the blocks of the experiment in Figure 3D. To determine if each individual showed significant adaptation to our formant perturbation, one-sample two-sided t-tests were run on each participant's trial-level data on this measure (n = 50) from the final adaptation block (block 6), to test for a significant difference from zero. This found significant adaptation (adaptation significantly greater than zero) in 13 out of 15 solo speech participants and 12 out of 15 synchronous speech participants. The remaining two participants in the solo speech group and one participant in the synchronous speech group showed adaptation responses that were not significantly different from zero. Two further participants in the synchronous speech group showed adaptation that was significantly lower than zero; this reflects a following response in which F1 and F2 were moved in the same direction as the feedback perturbation. A planned chi-squared test to compare the frequencies of adapters, non-adapters and followers between groups could not be performed due to too small expected frequencies.

To test the significance of adaptation at the group level, two-sided one-sample t-tests were used on F1 and F2 production changes from block 2 to block 6. These changes can be seen for each group in Figure 3C. For the solo speech group, these t-tests found a significant decrease in F1 ($t(14) = -5.56$, $p < .001$) but a significant increase in F2 ($t(14) = 6.74$, $p < .001$). Thus, as expected, the upwards perturbation of F1 and downwards perturbation of F2 drove significant compensatory changes in the produced F1 and F2 of participants in the solo speech group. For the synchronous speech group, these t-tests found a significant increase in F2 ($t(14) = 4.19$, $p < .001$), but no significant change in F1 ($t(14) = -1.95$, $p = .072$).

To directly compare adaptation between groups, a linear mixed modelling (LMM) analysis was run on our measure of adaptation using the lmerTest package in R (Kuznetsova et al., 2017). A random intercept model was run on adaptation values during blocks 3 to 7, with fixed effects of block and group, and random intercepts of participant and sentence. Random slopes were not included since these resulted in failures of model convergence. Both an additive model (in which block and group had additive effects) and an interactive model (in which block and group had interactive effects) were tested; a likelihood ratio test found that the interactive model did not provide a better fit to the data: $\chi^2(4) = 2.61$, $p = 0.625$. The additive model found no significant effect of group ($\beta = -0.457$, $t(28) = -0.131$, $p = .897$). Therefore, our hypothesis that the synchronous speech group would show either significantly reduced or enhanced adaptation was not supported. Follow-up contrasts using estimated marginal means found that both groups showed a significant difference in adaptation between block 3 and all other blocks, and between block 7 and all other blocks ($p < .002$ in all cases, using the Tukey method for adjusting for multiple comparisons). There were however no significant differences between blocks 4 and 5, 4 and 6, or 5 and 6. Thus, both groups showed an initial increase in adaptation with experience of altered feedback, which then remained stable until the removal of the feedback perturbation (see Figure 3D).

Exploratory Analyses

Although there were no significant differences in average adaptation at the group level, it can be seen in Figure 3A that there appears to be much greater between-individual variability in adaptation within the synchronous speech group versus the solo speech group. A Levene's test for equality of variances performed on participants' average adaptation in block 6 confirmed that there was a significant

difference in variability between the groups ($F(1,28) = 4.89$, $p = .036$). This suggests that synchronous speech may have affected adaptation responses in a participant-specific manner.

To explore what factors may underlie this increased variability, further exploratory analyses were conducted to investigate vocal convergence effects and their potential relationship to adaptation. Changes in F1, F2 and F0 from block 1 to block 2 allow for examination of acoustic-phonetic changes in speech productions related to performance of the synchronous speech task, before the feedback perturbation was introduced. These changes are plotted in Figure 4. LMM analyses were run for each measure separately to test for interactions between group and block, with the hypothesis that the effect of block should be significantly greater for the synchronous speech group compared to the solo speech group (who performed the same solo reading task across the two blocks). Significant group by block interactions were found for F2 ($\beta = 7.53$, $t(27.92) = 2.43$, $p = .022$) and F0 ($\beta = 10.72$, $t(29.19) = 10.65$, $p < .001$), with the synchronous speech group showing greater (upward) changes from block 1 to block 2 compared to the solo speech group. No significant interaction was found for F1 ($\beta = 6.75$, $t(28.03) = 1.956$, $p = .061$). To determine the significance of these changes, further LMM analyses were run on F0, F1 and F2 change (from block 1 to block 2), with the intercept suppressed (set to zero). This found that the changes in F2 ($\beta = 10.17$, $t(29.02) = 4.59$, $p < .001$) and F0 ($\beta = 10.69$, $t(28.52) = 4.24$, $p < .001$) were significantly greater than zero in the synchronous speech group. Neither changes in F1 in the synchronous speech group nor changes in any of the measures in the solo speech group were significant. As can be seen in Figure 5B and C, the average F2 and F0 of the accompanist voice stimuli were higher than the baseline block 1 averages of all participants in the synchronous speech group,

except for one (whose F2 was slightly higher than the accompanist). Thus, the synchronous speech group showed evidence of convergence to the accompanist's voice in the form of upward shifts in their F2 and F0 during block 2.

To consider how convergence and adaptation may relate to one another, we looked at whether the required formant changes for convergence and adaptation were in the same or different directions for each participant in the synchronous speech group. To do this, a convergence-adaptation congruency measure was obtained using an identical calculation to that used for our measure of adaptation, but substituting the participant's formant changes from block 2 to block 6 with the difference between the participant's formants at block 2 and the accompanist's formants. This measure thus quantifies the extent to which the formant changes required for the participant to converge to the accompanist voice lay on the same vector as perfect compensation for the subsequent formant perturbation. Individual vectors of convergence are plotted in Figure 5D. A Pearson's correlation analysis found that this convergence-adaptation congruency measure was correlated with adaptation responses within individuals (r = 0.531, $t(13)$ = 2.259, p = .042). That is, greater adaptation was found when the direction required for convergence to the accompanist voice was more similar to the direction required for adaptation.

It is of interest to note that for all but one of these participants, the formant changes required for convergence were broadly in the same direction as adaptation (i.e. all convergence-adaptation congruency values were positive). However, for one participant these were clearly in the opposite direction, resulting in a negative value. Interestingly, this participant also showed a large following response to the formant perturbation. This exploratory analysis therefore provides preliminary support for the

idea that the effect of synchronous speech on measured adaptation responses may depend on the congruency between the direction of change required for convergence and that required for adaptation.

To assess whether synchronisation success was related to individual variability in either vocal convergence or adaptation, Pearson's correlations were run using the asynchrony scores calculated from block 2 (see *Synchronisation Success*). These found no significant correlations between asynchrony scores and either adaptation (averaged across block 6) or convergent changes in F0, F1 or F2 from block 1 to 2.

Experiment 2

The results of Experiment 1 generally did not confirm our hypothesis that synchronous speech would result in an increased or decreased magnitude of adaptation. Although production changes in F1 failed to reach significance for this group, when considering F1 and F2 changes together in our adaptation measure, the average magnitude of adaptation was not significantly different within the solo speech and synchronous speech groups. The results do however suggest that synchronous speech resulted in increased between-participant variability in the magnitude and direction of the measured adaptation response. Exploratory analyses suggested the hypothesis that the effect of synchronous speech on adaptation may depend on the acoustic-phonetic properties of the accompanist voice; specifically, on whether the formant changes required for convergence aligned with those required for adaptation. To test this idea, we ran a second experiment that investigated the effect of metronome-timed speech on adaptation. This preserves the temporal synchronisation aspect of the synchronous speech task, and the presence of an external auditory cue during speech, but crucially does not involve exposure to

another voice. If the increased between-individual variability in adaptation during synchronous speech was specifically related to perception of the accompanist voice, we would predict that participants in a metronome-timed speech condition should show similar adaptation to the solo speech group from Experiment 1, both in terms of the group average adaptation response and the level of between-participant variability.

Methods

Participants

Fifteen female participants (*mean age* = 21.24 years, *sd* = 1.79) were recruited and tested with the metronome-timed condition. As in Experiment 1, all participants were native speakers of British English from the South-East of England, with no reported history of hearing loss or history of speech, language or reading disorders. The data from this new sample was compared to existing data from the sample of 15 participants tested with the solo speech condition in Experiment 1.

Procedure

The general procedure, apparatus and formant perturbation used in this experiment were identical to the synchronous speech condition from Experiment 1; however, for this condition, from block 2 onwards the participants were instead instructed to synchronise their speech with a metronome-beat that was played through the headphones. This metronome beat had a rate of 186 beats per minute; this value was chosen to approximately match the speaking rate (quantified as the average number of syllables per second) of the accompanist speech recordings from Experiment 1. Participants were instructed to produce one speech syllable per beat; they were first given practice with a slower beat and then the faster beat before

starting block 2. The metronome beat was played continuously from the start to the end of the trial, with each trial lasting 11 seconds. Participants had choice over when to begin speaking within the trial (to ensure they started in time with the beat); the trial length was therefore slightly increased from that used in Experiment 1 in order to allow for differences in when participants started speaking (i.e. to ensure their sentence productions were not prematurely cut off).

Hypotheses

We predicted that participants in this metronome-timed speech condition would show significant adaptation to the formant perturbation. Furthermore, we predicted that this adaptation would look similar to that observed in the solo speech condition from Experiment 1, both in terms of the magnitude of the group average adaptation response, and the level of between-individual variability in adaptation. Lastly, we predicted that metronome-speech itself would not cause any significant changes to produced formants before the formant perturbation was introduced; that is, there would be no significant difference in F1 and F2 values between block 1 (solo reading) and block 2 (metronome-timed speech).

Results

Individual adaptation vectors are illustrated in Figure 6A; individual after-effects of adaptation are illustrated as vectors in Figure 6B. Adaptation responses quantified as the component of each participant's formant changes that directly opposed the perturbation are illustrated in Figure 6D. Significant adaptation (adaptation significantly greater than zero) was found in 11 out of 15 participants in the metronome-timed speech group, with the remaining four showing no significant difference from zero (two-sided one-sample t-tests).

Changes in produced F1 and F2 from baseline are illustrated in Figure 6C. At the group level, two-sided one-sample t-tests found that the metronome-timed speech group showed a significant increase in F2 ($t(14) = 6.28$, $p < .001$), but no significant change in F1 ($t(14) = -1.35$, $p = .198$). This matches the pattern of results for the synchronous speech group in Experiment 1.

To directly compare adaptation in the metronome-timed speech group to the solo speech group, a linear mixed modelling (LMM) analysis was again run on our adaptation measure, modelling fixed effects of block (3-7) and group, and random intercepts of participant and sentence. A likelihood ratio test found that an interactive model did not provide a better fit to the data than an additive model: $\chi^2(4) = 2.649$, $p = .618$. The additive model found no significant effect of group ($\beta = 0.822$, $t(27.99) = 0.308$, $p = .761$). This confirms our hypothesis that the average magnitude of adaptation would not differ significantly between metronome-timed and solo speech groups. Follow-up contrasts using estimated marginal means found that both groups showed a significant difference in adaptation between block 3 and each of blocks 4, 5 and 6, and between block 7 and blocks 4, 5 and 6 ($p < .002$ in all cases, using the Tukey method for adjusting for multiple comparisons). There were however no significant differences between blocks 3 and 7, 4 and 5, 4 and 6, or 5 and 6.

To compare the level of between-participant variability in adaptation between the two groups, an F-test was used to compare the variances of the two samples on our adaptation measure (since a Shapiro-Wilk test indicated that the data were normally distributed). This found significantly greater between-participant variability in adaptation in the metronome-timed speech group compared to the solo speech group: $F(14,14) = 0.294$, $p = .029$. This does not confirm our hypothesis that the level

of between-participant variability in adaptation would not significantly differ between the metronome-timed and solo speech groups.

Lastly, to further explore the impact of metronome-timed speech on the acoustic-phonetic properties of participants' speech before the formant perturbation was applied, we looked at changes in F0, F1 and F2 from block 1 to block 2. These changes are plotted in Figure 7. LMM analyses were run for each measure separately to test for interactions between group (metronome-timed speech versus solo speech) and block (1 versus 2). Significant group by block interactions were found for F0 (β = -3.39, t(2919) = -3.33, p < .001), for F1 (β = 7.84, t(2913.01) = 4.43, p < .001) and for F2 (β = 12.30, t(2913) = 3.18, p = .001), indicating that the metronome-timed group showed greater changes between blocks 1 and 2 than the solo speech group. To determine the significance of these changes, further LMM analyses were run on F0, F1 and F2 change (from block 1 to block 2), with the intercept suppressed (set to zero). This found that neither the changes in F0 (β = -3.42, t(28.16) = -1.29, p = .207) nor the changes in F1 (β = 5.03, t(33.01) = 1.79, p = .082) were significantly different from zero in the metronome-timed speech group. In contrast, F2 changes were significantly different from zero in this group (β = 14.89, t(52.84) = 5.16, p < .001). Thus, contrary to our hypothesis, metronome-timed speech resulted in significant increases in F2 before the feedback perturbation was introduced.

Overall Discussion

This study investigated speech motor adaptation to a formant perturbation during fluid sentence production in three conditions involving different speech tasks. In Experiment 1, we found similar levels of adaptation during a synchronous speech

task and a solo speech task when considering the group average adaptation response. Synchronous speech thus did not appear to have any large systematic effects on the average magnitude of the measured speech motor adaptation response across the group. However, between-individual variability in adaptation responses was significantly increased in the synchronous speech condition, suggesting that the task may have had participant-specific effects on measured formant changes. An exploratory correlation analysis suggested the hypothesis that the effect of synchronous speech may depend on the congruency between the formant changes required for phonetic convergence to the accompanist voice, and those required for adaptation. As a preliminary test of this hypothesis, in Experiment 2 we measured adaptation during metronome-timed speech, a task that similarly requires synchronisation of speech with an external input, but crucially in the absence of another voice. Unexpectedly, adaptation in this condition appeared similar to that shown during synchronous speech; although the average magnitude of adaptation did not differ from the solo speech group, again the metronome-timed group demonstrated significantly greater between-participant variability in adaptation responses. Overall therefore, these findings suggest that the average magnitude of the adaptation response at the group level can be robust across more complex speaking contexts. However, the observed increase in between-participant variability may signal the influence of uncontrolled factors introduced by the tasks, that could have prevented the emergence of systematic effects across participants.

Increased between-individual variability in adaptation behaviour does not appear to be simply an inevitable consequence of engaging in a secondary task. A dual-tasking study by Lametti et al., (2020) found a similar average magnitude and level of between-participant variability of speech adaptation responses both in the

presence and absence of performance of a concurrent visuomotor adaptation task. Furthermore, in the current study, an exploratory correlation analysis suggested that the increased variability in the synchronous speech group may be meaningfully related to individual differences in vocal convergence processes. Across the group, we found evidence of convergence in participants' F0 and F2 towards those of the accompanist voice, replicating previous work (for F0; Bradshaw & McGettigan, 2021). Further, the magnitude of adaptation during synchronous speech was found to be positively correlated with the extent to which the formant changes required for convergence and adaptation were in the same direction within participants. That is, greater adaptation was shown when the direction of formant change required for convergence to the accompanist voice more closely matched the direction for adaptation (i.e. a downward shift in F1 but an upward shift in F2). This exploratory analysis suggests the hypothesis that individual differences in measured adaptation responses during synchronous speech could be due to acoustic-phonetic properties of the accompanist voice, and their relation to the voice of each individual participant.

It is however difficult to determine the nature of the potential interaction between convergence-related and adaptation-related processes based on this data alone. That is, the increased variability in the measured adaptation response could reflect effects of convergence on the process of adaptation itself; or alternatively, it could reflect the simple linear sum of convergence-related formant changes with those arising from an (itself unaffected) sensorimotor adaptation process. One potentially informative feature of our experimental design concerns the inclusion of two baseline blocks at the start of the experiment; the first with solo speech and the second with synchronous/metronome-timed speech *before* the feedback alteration was introduced. Crucially, adaptation was always measured relative to this second

baseline block, meaning that at least some of the changes in formants related to convergence to the accompanist voice were controlled for. Whether convergence continued to drive changes in speech productions into the altered feedback section of the experiment however is debatable. Previous work suggests that vocal convergence can have a rapid onset and a stable trajectory across a task; this has been reported specifically for F0 convergence during synchronous speech (Bradshaw & McGettigan, 2021), as well as for other measures such as vowel duration and spectral distributions during other tasks (Aubanel & Nguyen, 2020; Delvaux & Soquet, 2007). However, convergence-related processes clearly continued to have some influence on compensatory formant changes measured relative to block 2, given the significant correlation with our measure of convergence-adaptation congruency; that is, the relationship between the participant's and the accompanist's voice formants still influenced the magnitude of compensatory formant changes even when accounting for initial formant changes induced by synchronous speech in the absence of altered feedback.

Interestingly, in this sample all but one of the participants had baseline formant frequencies that resulted in agreement between the direction of change required for convergence and adaptation. The one participant for whom these measures disagreed was also exceptional in showing a large following response, moving their formant frequencies in the same direction as the formant perturbation (and thus in the direction towards the accompanist voice). This suggests an ongoing convergence response that thus interfered with compensatory formant changes to the feedback perturbation. It would be of interest to follow up this exploratory finding with an experiment that explicitly manipulated the congruency between the formant changes required for convergence and adaptation by modifying the acoustic-

phonetic properties of the accompanist voice. Such a design may allow for demonstration of more systematic effects of synchronous speech on adaptation responses across a group, and yield insights into the underlying relationship between convergence and adaptation processes. For example, adaptation could be measured in the same individual firstly during solo speech, and then again during synchronous speech with an accompanist voice manipulated to pull the speaker's formants in a direction opposite to that of adaptation. If the effects of adaptation and convergence linearly sum, we would expect adaptation to be significantly greater during solo speech compared to this synchronous speech condition; conversely, if there is an interaction and adaptation is prioritised, we would expect measured adaptation to not significantly differ across these two conditions.

Contrary to our predictions, Experiment 2 found a similar increase in between-participant variability in measured adaptation responses during performance of a speech synchronisation task that does not involve another voice (metronome-timed speech). This condition was also unexpectedly associated with phonetic changes prior to introduction of the formant perturbation, in the form of upward shifts in produced F2. These may be related to changes in prosodic speech features induced by the metronome task; namely, changes in speech rate and stress. There is long-standing evidence that such changes can affect the spectral quality of vowels, typically mediated by effects on vowel duration (i.e. shortening of vowels at faster speech rates) and on coarticulation (with some evidence that coarticulation effects increase at higher speech rates) (Gay, 1978; Pitermann, 2000; Recasens, 2015; Weismer & Berry, 2003). Such studies have often highlighted F2 as particularly sensitive to effects of speaking rate and stress (Gay, 1978; Recasens, 2015; Weismer & Berry, 2003), concurring with the current finding of effects of metronome-

timed speech on F2 only. Furthermore, these effects of speech rate on vowel duration and vowel quality can often be subject-specific (Recasens, 2015; Weismer & Berry, 2003). This suggests the possibility that some of the increased variability in F1 and F2 changes during this condition may have arisen from between-participant variability in the effects of metronome speech itself on formants. That is, it is again possible that the observed increased variability does not reflect the influence of a synchronisation task on the process of adaptation itself, but merely the summation of adaptation-related and metronome speech-related formant changes. As previously noted, adaptation was measured relative to the second baseline block in an attempt to control for any effects of the metronome-timed task on speech productions in the absence of perturbation; however, it is still possible that these changes continued to occur during the altered feedback phase. Indeed, these formant changes relating to changes in speech rate and stress may mask the 'pure' effects of the speech timing synchronisation process itself; for example, by inducing incidental increases in F2 that were in fact not related to adaptation to the downwards F2 perturbation experienced. This may explain the high number of participants apparently showing adaptation on F2 only (see Figure 6). Replication of this experiment using a different formant perturbation (e.g. an upwards shift in F2 or no perturbation of F2) would be informative with regards to this interpretation.

Investigating sensorimotor control during rhythmic speech behaviours is of clinical relevance to the study of developmental stuttering. Stuttering is a disorder affecting speech motor control that involves frequent dysfluencies during the production of speech, in the form of sound repetitions and prolongations, and pauses in which a speech sequence fails to be initiated (known as blocks). It has long been observed that both synchronous and metronome-timed speech can result in a temporary

enhancement of speech fluency in people who stutter (Bloodstein & Ratner, 2008; Kalinowski & Saltuklaroglu, 2003), suggesting they share important underlying properties during speech motor control. It has also been demonstrated that people who stutter show significantly reduced speech motor adaptation to perturbations of auditory feedback (for a review, see Bradshaw et al., 2021). It would thus be of interest to investigate whether synchronous or metronome-timed speech may facilitate an enhancement of the adaptation response in people who stutter. Interestingly, a recent study by Frankford et al., (2022) investigated the effects of metronome-timed speech on compensation to unexpected (i.e. randomly occurring) perturbations of F1 in adults who do and do not stutter. They found a significant group by condition interaction; while significant compensation was found for a metronome-timed speech but not a normal speech condition in adults who stutter, the reverse was found for adults who do not stutter. Compensation for unexpected auditory feedback perturbations is typically attributed to the 'online' operation of a feedback control system; this is distinguished from the adaptation to sustained feedback perturbations under investigation in the current paper, which is attributed to offline updating of forward models in the feedforward control system (Guenther, 2016). Comparing the effects of external pacing conditions (such as synchronous and metronome-timed speech) on adaptation versus compensation responses in people who stutter could thus potentially be informative for deciphering the nature of the disruption to speech sensorimotor control in this speech disorder.

Overall, the findings of this study demonstrate that while the average magnitude of measured speech motor adaptation responses can be robust at the group level during performance of more complex speaking tasks, differences in the underlying variability of individual adaptation responses suggest the additional influence of task-

related factors such as vocal convergence and the coordination of speech timing with external stimuli and other voices. This suggests that speech motor behaviour can be shaped by a variety of external factors such as the social, perceptual and task-based aspects of the speaking context, rather than solely being driven by the simple calculation of sensory prediction errors in the self voice. It would be informative therefore to consider how current neurocognitive models of speech motor control might incorporate influences of these external factors (Houde & Nagarajan, 2011; Parrell, Lammert, et al., 2019; Parrell, Ramanarayanan, et al., 2019; Tourville & Guenther, 2011). This would enable such models to account for the dynamics of sensorimotor control of speech in more complex speaking contexts. Further work is needed to determine the mechanism by which these task-related factors may affect the measured adaptation response; specifically, whether this reflects a linear summation of the independent effects of adaptation versus synchronisation-/convergence-related formant changes, or whether such task-related processes affect the process of speech motor learning itself. This will provide further insights into the relationships between speech motor adaptation and vocal convergence, and may contribute to our understanding of the sensorimotor mechanisms behind dysfluent speech in stuttering.

Constraints on Generality Statement

In our experiments, we tested samples of female speakers of British English, with a standard Southern British English accent. We believe the results should generalise across both males and females as well as across different accents, as long as the sex/gender and accent of the participant and the accompanist are matched to each other. We would expect the findings to also generalise across other languages. Our experiments also used variable sentence stimuli, and so we would expect the

findings to generalise to other sentences or short pieces of text not tested here. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

References

Aubanel, V., & Nguyen, N. (2020). Speaking to a common tune: Between-speaker convergence in voice fundamental frequency in a joint speech production task. *PLOS ONE*, *15*(5). https://doi.org/10.1371/journal.pone.0232209

Audacity Team. (2021). *Audacity(R): Free Audio Editor and Recorder*. https://audacityteam.org/

Bloodstein, O., & Ratner, N. (2008). *A Handbook on Stuttering*. Clifton Park.

Boersma, P. (1993). Accurate Short-Term Analysis Of The Fundamental Frequency And The Harmonics-To-Noise Ratio Of A Sampled Sound. *Proceedings of the Institute of Phonetic Sciences*, *17*.

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer*.

Bourguignon, N. J., Baum, S. R., & Shiller, D. M. (2016). Please Say What This Word Is-Vowel-Extrinsic Normalization in the Sensorimotor Control of Speech. *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-HUMAN PERCEPTION AND PERFORMANCE*, *42*(7), 1039–1047. https://doi.org/10.1037/xhp0000209

Bradshaw, A. R., Lametti, D. R., & McGettigan, C. (2021). The Role of Sensory Feedback in Developmental Stuttering: A Review. *Neurobiology of Language*, *2*(2), 1–27. https://doi.org/10.1162/nol_a_00036

Bradshaw, A. R., & McGettigan, C. (2021). Convergence in voice fundamental frequency during synchronous speech. *PLOS ONE*, *16*(10), e0258747. https://doi.org/10.1371/journal.pone.0258747

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA*, *103*(6), 3153–3161. https://doi.org/10.1121/1.423073

Cai, S. (2015). *Audapter*.

Cai, S., Boucek, M., Ghosh, S., Guenther, F., & Perkell, JS. (2008). A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. *Proceedings of the 8th Intl. Seminar on Speech Production*, 65–68.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, *37*(1), 16–28. https://doi.org/10.1016/j.wocn.2008.08.003

Cummins, F. (2018). Joint speech as an object of empirical inquiry. *Material Religion*, *14*(3), 417–419. https://doi.org/10.1080/17432200.2018.1485344

Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *PHONETICA*, *64*(2–3), 145–173. https://doi.org/10.1159/000107914

Ellis, D. (2003). *Dynamic Time Warp (DTW) in Matlab*. http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/.

Frankford, S. A., Cai, S., Nieto-Castañón, A., & Guenther, F. H. (2022). Auditory feedback control in adults who stutter during metronome-paced speech II. Formant Perturbation. *Journal of Fluency Disorders*, 105928. https://doi.org/10.1016/j.jfludis.2022.105928

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *The Journal of the Acoustical Society of America*, *63*(1), 223–230. https://doi.org/10.1121/1.381717

Giles, H., Coupland, N., & Coupland, J. (1991). 1. Accommodation theory: Communication, context, and consequence. In *Contexts of Accommodation: Developments in Applied Sociolinguistics* (pp. 1–68). Cambridge University Press.

Guenther, F. H. (2016). *Neural Control of Speech*. The MIT Press.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*(3), 280–301. https://doi.org/10.1016/j.bandl.2005.06.001

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213–1216. https://doi.org/10.1126/science.279.5354.1213

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, *OCTOBER*. https://doi.org/10.3389/fnhum.2011.00082

IEEE Subcommittee on Subjective Measurements. (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 227–246.

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. https://doi.org/10.1523/JNEUROSCI.4075-15.2016

Kalinowski, J., & Saltuklaroglu, T. (2003). Choral speech: The amelioration of stuttering via imitation and the mirror neuronal system. *Neuroscience & Biobehavioral Reviews*, *27*(4), 339–347. https://doi.org/10.1016/S0149-7634(03)00063-0

Kim, K. S., Wang, H., & Max, L. (2020). It's About Time: Minimizing Hardware and Software Latencies in Speech Research With Real-Time Auditory Feedback. *Journal of Speech, Language, and Hearing Research*, *63*(8), 2522–2534. https://doi.org/10.1044/2020_JSLHR-19-00419

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *JOURNAL OF STATISTICAL SOFTWARE*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lametti, D. R., Krol, S. A., Shiller, D. M., & Ostry, D. J. (2014). Brief Periods of Auditory Perceptual Training Can Determine the Sensory Targets of Speech Motor Learning. *Psychological Science*. https://doi.org/10.1177/0956797614529978

Lametti, D. R., Quek, M. Y. M., Prescott, C. B., Brittain, J.-S., & Watkins, K. E. (2020). The perils of learning to move while speaking: One-sided interference between speech and visuomotor adaptation. *Psychonomic Bulletin & Review*, *27*(3), 544–552. https://doi.org/10.3758/s13423-020-01725-8

Lametti, D. R., Smith, H. J., Watkins, K. E., & Shiller, D. M. (2018). Robust Sensorimotor Learning during Variable Sentence-Level Speech. *Current Biology*, *28*(19), 3106-3113.e2. https://doi.org/10.1016/j.cub.2018.07.030

Max, L., & Maffett, D. G. (2015). Feedback delays eliminate auditory-motor learning in speech production. *NEUROSCIENCE LETTERS*, *591*, 25–29. https://doi.org/10.1016/j.neulet.2015.02.012

Niziolek, C. A., & Guenther, F. H. (2013). Vowel Category Boundaries Enhance Cortical and Behavioral Responses to Speech Feedback Alterations. *Journal of Neuroscience*, *33*(29), 12090–12098. https://doi.org/10.1523/JNEUROSCI.1008-13.2013

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, and Psychophysics*, *79*(2), 637–659. https://doi.org/10.3758/s13414-016-1226-0

Parrell, B., Lammert, A. C., Ciccarelli, G., & Quatieri, T. F. (2019). Current models of speech motor control: A control-theoretic overview of architectures and properties. *JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA*, *145*(3), 1456–1481. https://doi.org/10.1121/1.5092807

Parrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. F. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLOS COMPUTATIONAL BIOLOGY*, *15*(9). https://doi.org/10.1371/journal.pcbi.1007321

Pitermann, M. (2000). Effect of speaking rate and contrastive stress on formant dynamics and vowel perception. *The Journal of the Acoustical Society of America*, *107*(6), 3425–3437. https://doi.org/10.1121/1.429413

Purcell, D. W., & Munhall, K. G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.2217714

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org

Recasens, D. (2015). The Effect of Stress and Speech Rate on Vowel Coarticulation in Catalan Vowel-Consonant-Vowel Sequences. *JOURNAL OF SPEECH*

LANGUAGE AND HEARING RESEARCH, *58*(5), 1407–1424. https://doi.org/10.1044/2015_JSLHR-S-14-0196

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L., & Nguyen, N. (2013). Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in Psychology*, *4*, 422. https://doi.org/10.3389/fpsyg.2013.00422

Sato, M., & Shiller, D. M. (2018). Auditory prediction during speaking and listening. *BRAIN AND LANGUAGE*, *187*, 92–103. https://doi.org/10.1016/j.bandl.2018.01.008

Shiller, D. M., Mitsuya, T., & Max, L. (2020). Exposure to Auditory Feedback Delay while Speaking Induces Perceptual Habituation but does not Mitigate the Disruptive Effect of Delay on Speech Auditory-motor Learning. *NEUROSCIENCE*, *446*, 213–224. https://doi.org/10.1016/j.neuroscience.2020.07.041

Shiller, D. M., & Rochon, M. L. (2014). Auditory-Perceptual Learning Improves Speech Motor Adaptation in Children. *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-HUMAN PERCEPTION AND PERFORMANCE*, *40*(4), 1308–1315. https://doi.org/10.1037/a0036660

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *LANGUAGE AND COGNITIVE PROCESSES*, *26*(7), 952–981. https://doi.org/10.1080/01690960903498424

Weismer, G., & Berry, J. (2003). Effects of speaking rate on second formant trajectories of selected vocalic nuclei. *The Journal of the Acoustical Society of America*, *113*(6), 3362–3378. https://doi.org/10.1121/1.1572142

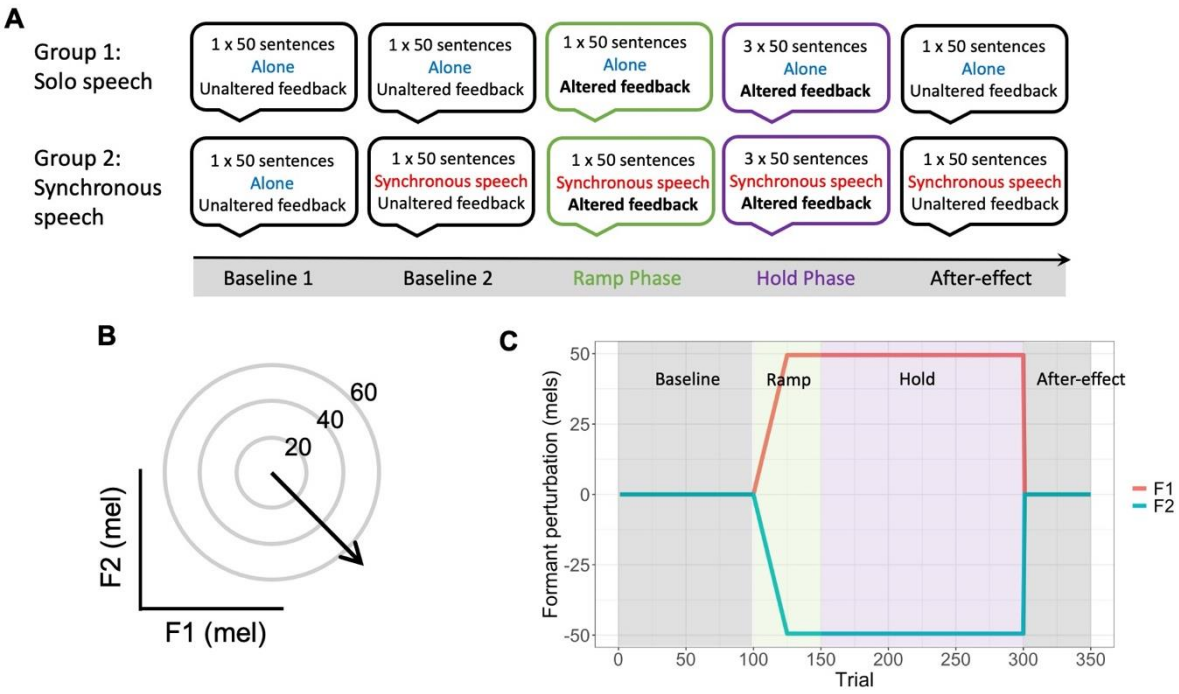

Figure 1: Experimental Procedure. *(A) Structure of the experiment for the two conditions: solo speech and synchronous speech. (B) Schematic illustration of the F1 up F2 down (F1+ F2-) perturbation employed (joint perturbation magnitude of 70 mels). (C) Schematic illustration of the timing and magnitude of the perturbation of F1 and F2 in mels. Shading indicates phase as labelled (baseline, ramp, hold and after-effect).*

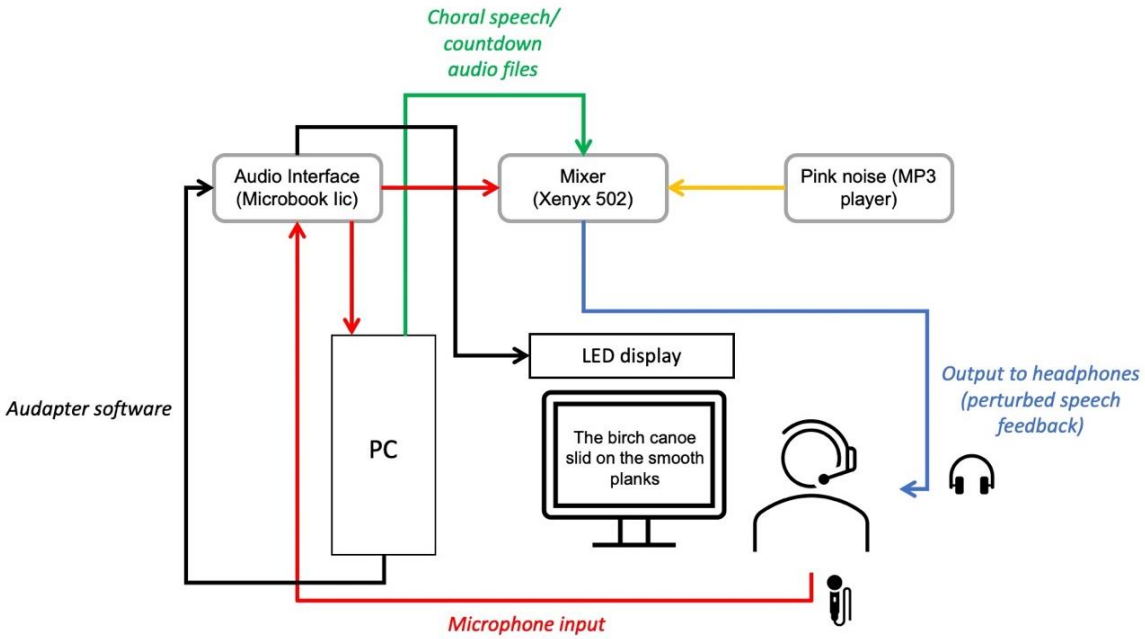

Figure 2: Experimental set-up for speech motor adaptation paradigm.

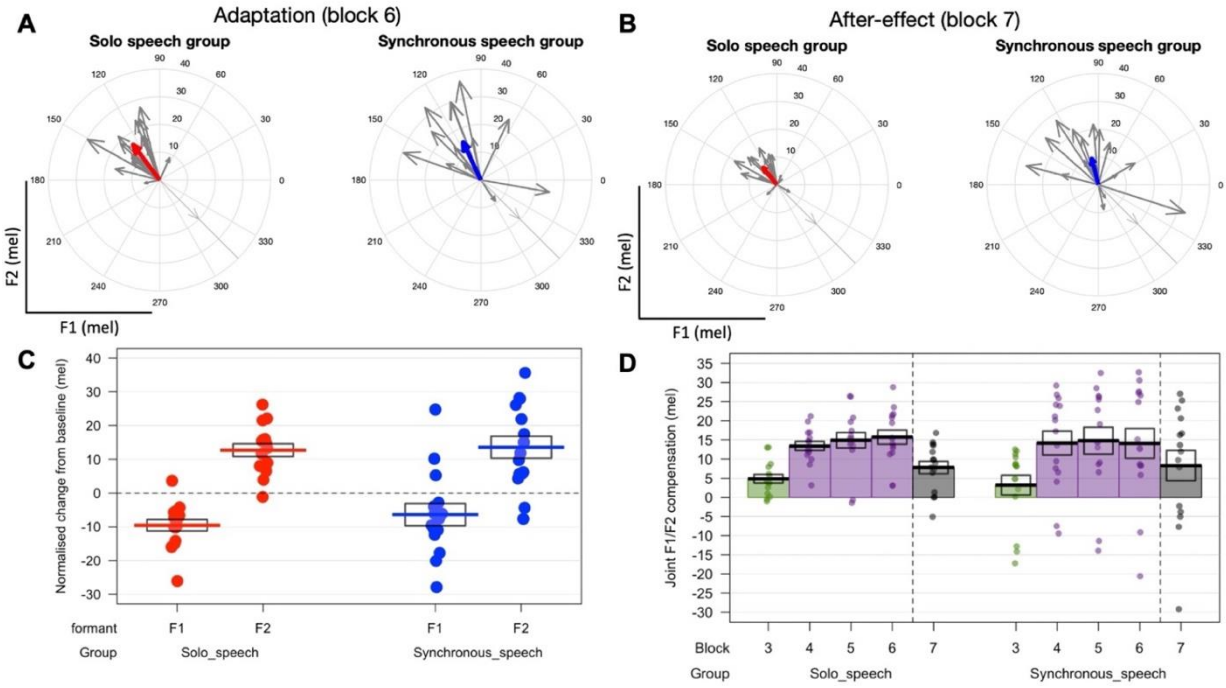

Figure 3: Sensorimotor Adaptation during solo speech and synchronous speech. *(A) Thin grey arrows indicate adaptation responses for each participant (in block 6), in the form of vectors in F1/F2 space. The group average responses are shown in thick arrows in red (light grey) and blue (dark grey) for solo and*

synchronous speech groups respectively. The pale grey arrow at 315 degrees indicates the feedback perturbation direction. (B) Equivalent data to A for the after-effects of adaptation (block 7) when the perturbation was removed. (C) Production change in mel for F1 and F2 from baseline block 2 to the final block of perturbed feedback (block 6) in the solo speech and synchronous speech groups. Dots indicate individual participant averages; thick lines indicate group means and boxes indicate standard errors. (D) Adaptation responses (quantified as the component of formant changes that directly opposed the perturbation direction) for blocks 3-7 in the solo speech and synchronous speech groups. Dots indicate individual participant data, thick lines indicate group means, and boxes indicate standard errors. Colour coding of bars indicates phase: green (light grey) shows the ramp phase (formant perturbation gradually increased), purple (mid-grey) shows the hold phase (perturbation held constant), and black (dark grey) shows the after-effect phase (perturbation removed). Dotted vertical lines indicate removal of the feedback perturbation for block 7.

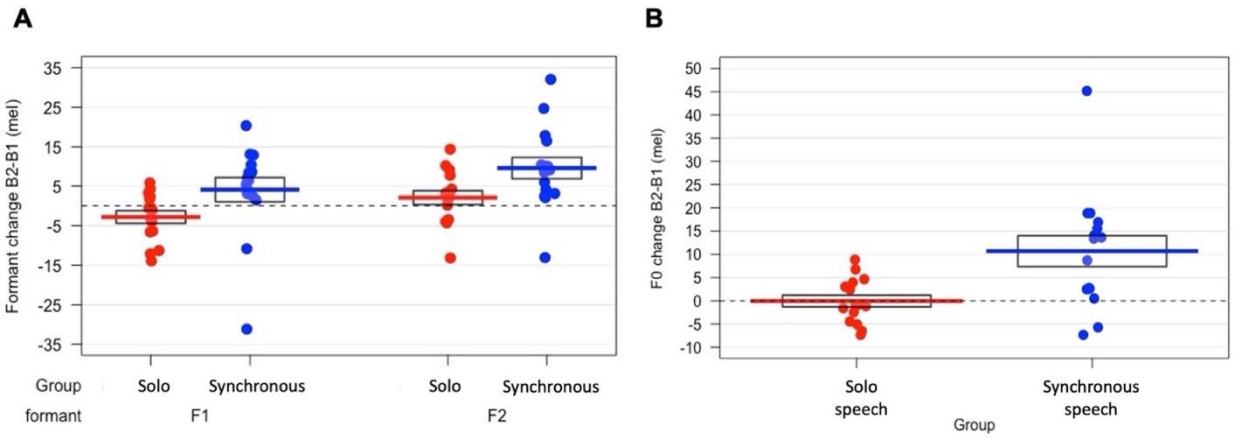

Figure 4: Vocal convergence in F1, F2 and F0. *(A) F1 and F2 changes from block 1 to block 2 in mels in the solo and synchronous speech groups. Dashed line indicates zero. (B) F0 changes from block 1 to block 2 in mels in the solo and synchronous speech groups. Dashed line indicates zero.*

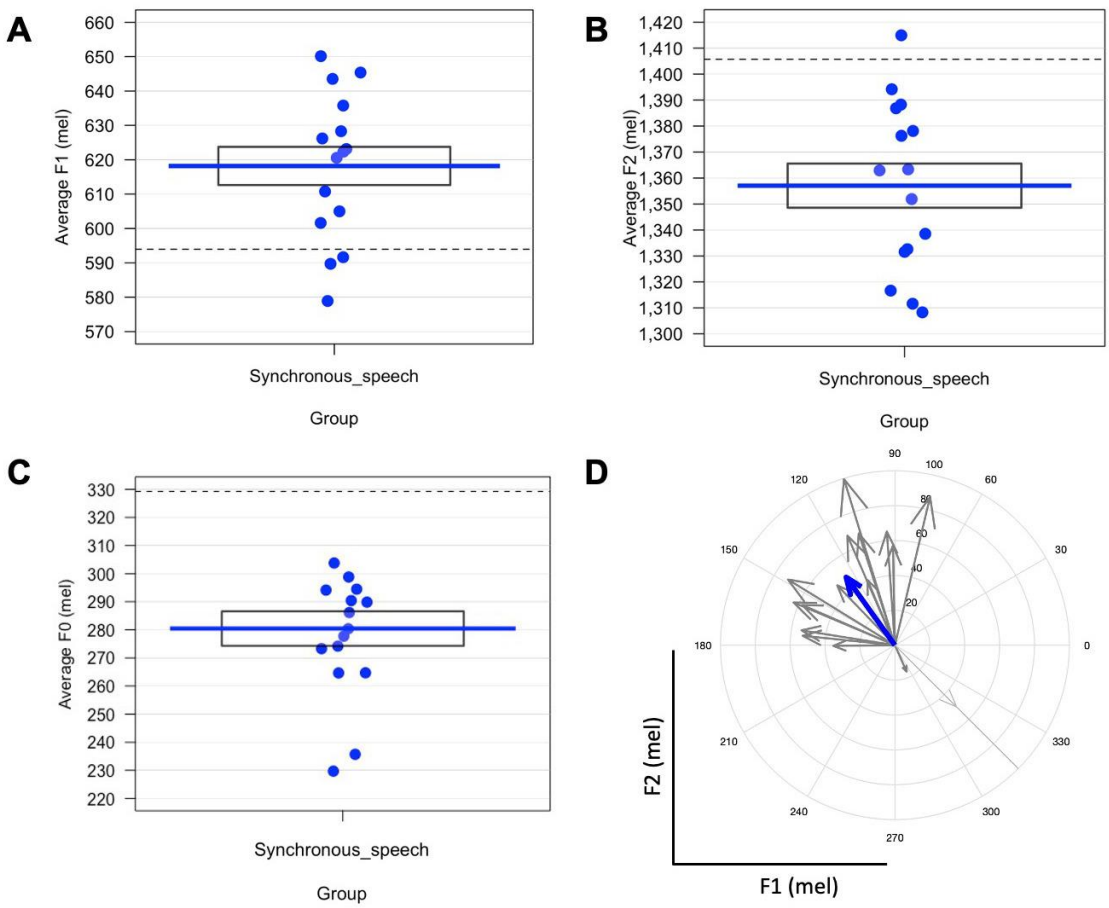

Figure 5: Distance from the accompanist voice. *(A) Average F1 values during block 1 for participants in the synchronous speech group. Dashed line indicates the average F1 of the accompanist voice. (B) Average F2 values during block 1 for participants in the synchronous speech group. Dashed line indicates the average F2 of the accompanist voice. (C) Average F0 values during block 1 for participants in the synchronous speech group. Dashed line indicates the average F0 of the accompanist voice. (D) Vectors of perfect convergence for the synchronous speech group, showing the direction of formant changes required for convergence to the accompanist voice. Thin grey arrows indicate individual participants, thick blue (dark grey) arrow indicates group average. The pale grey arrow at 315 degrees indicates the feedback perturbation direction subsequently experienced in the experiment.*

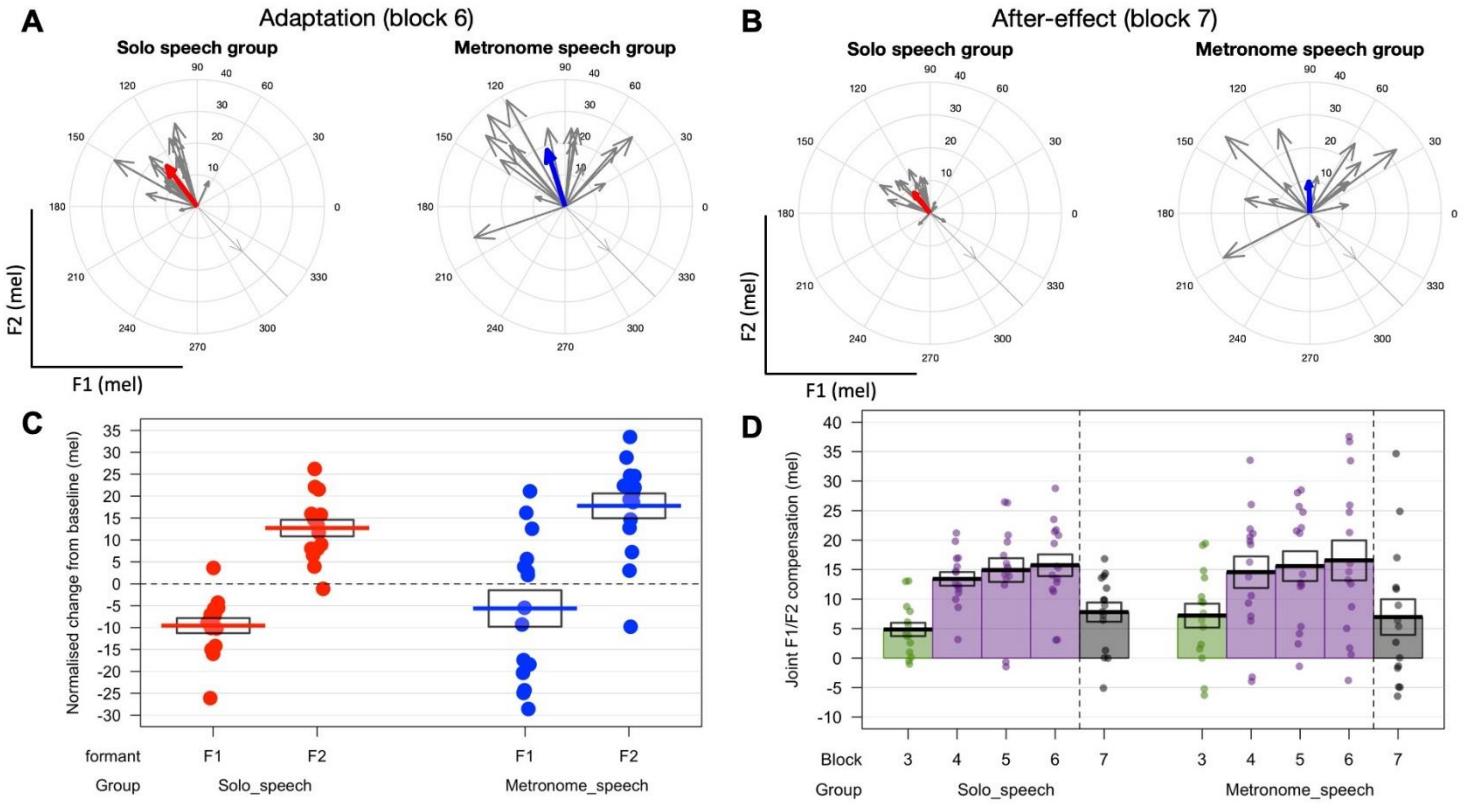

Figure 6: Sensorimotor Adaptation during solo speech and metronome-timed speech. *(A) Thin grey arrows indicate adaptation responses for each participant (in block 6), in the form of vectors in F1/F2 space. The group average responses are shown in thick arrows in red (light grey) and blue (dark grey) for solo and metronome-timed speech groups respectively. The pale grey arrow at 315 degrees indicates the feedback perturbation direction. (B) Equivalent data to A for the after-effects of adaptation (block 7) when the perturbation was removed. (C) Production change in mel for F1 and F2 from baseline block 2 to the final block of perturbed feedback (block 6) in the solo speech and metronome-timed speech groups. Dots indicate individual participant averages, thick lines indicate group means and boxes indicate standard errors. (D) Adaptation responses for blocks 3-7 in the solo speech and metronome-timed speech groups. Dots indicate individual participant data, thick lines indicate group means, and boxes indicate standard errors. Colour coding of bars indicates phase: green (light grey) shows the ramp phase (formant perturbation*

gradually increased), purple (mid-grey) shows the hold phase (perturbation held constant), and black (dark grey) shows the after-effect phase (perturbation removed). Dotted vertical lines indicate removal of the feedback perturbation for block 7.

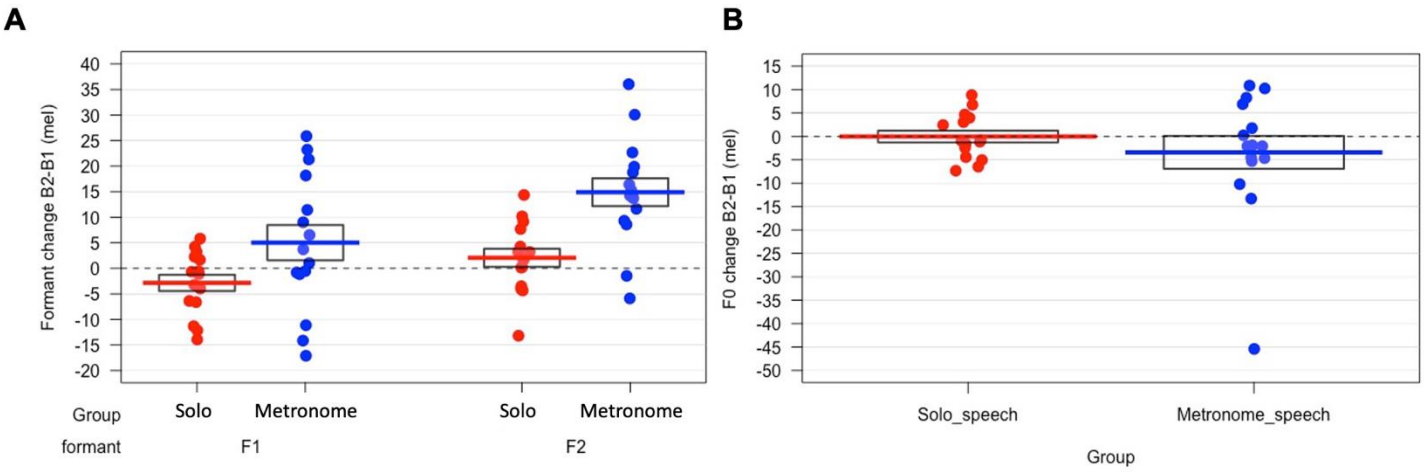

Figure 7: Changes in F1 and F2 from block 1 to block 2. *(A) F1 and F2 changes from block 1 to block 2 in mels in the solo and metronome-timed speech groups. Dashed line indicates zero. (B) F0 changes from block 1 to block 2 in mels in the solo and metronome-timed speech groups. Dashed line indicates zero.*