Discourse Processes
A Multidisciplinary Journal
The Official Journal of The Society for Text and Discourse
DP
Routledge

Discourse Processes

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hdsp20

# Multimodal Coordination of Sound and Movement in Music and Speech

Camila Alviar , Rick Dale , Akeiylah Dewitt & Christopher Kello

**To cite this article:** Camila Alviar , Rick Dale , Akeiylah Dewitt & Christopher Kello (2020) Multimodal Coordination of Sound and Movement in Music and Speech, Discourse Processes, 57:8, 682-702, DOI: 10.1080/0163853X.2020.1768500

**To link to this article:** https://doi.org/10.1080/0163853X.2020.1768500

Published online: 18 Jun 2020.

Submit your article to this journal

Article views: 71

View related articles

View Crossmark data

DISCOURSE PROCESSES
2020, VOL. 57, NO. 8, 682–702
https://doi.org/10.1080/0163853X.2020.1768500

ARTICLES

# Multimodal Coordination of Sound and Movement in Music and Speech

Camila Alviar[a], Rick Dale[b], Akeiylah Dewitt[a], and Christopher Kello[a]

[a]Cognitive and Information Sciences, University of California, Merced; [b]Department of Communication, University of California, Los Angeles

**ABSTRACT**

Speech and music emerge from a spectrum of nested motor and perceptual coordination patterns across timescales of brief movements to actions. Intuitively, this nested clustering in movements should be reflected in sound. We examined similarities and differences in multimodal, multiscale coordination of speech and music using two complementary measures: We computed spectra for envelopes of acoustic amplitudes and motion amplitudes and correlated spectral powers across modalities as a function of frequency. We also correlated smoothed envelopes and examined peaks in their cross-correlation functions. YouTube videos of five different modes of speaking and five different types of music were analyzed. Speech performances yielded stronger, more reliable relationships between sound and movement compared with music. Interestingly, a cappella singing patterned more with music, and improvisational jazz piano patterned more with speech. Results suggest that nested temporal structures in sound and movement are coordinated as a function of communicative aspects of performance.

## Introduction

Whenever we see someone speaking or playing an instrument, the sound we hear is experienced as integrated with the movement we see. In fact, movements of the lungs, vocal folds, tongue, and lips causally produce the speech we hear, and movements of the fingers, arms, lungs, torso, and even feet causally produce the music we hear. These movements may seem somewhat isolated and subtle at times, even invisible to the perceiver, but the musculoskeletal system works in concert to produce many kinds of movements (Bernstein, 1967). Posture is dynamically adjusted and poise maintained with visible body movements, even when targeted movements may be difficult to perceive. The overarching aims of our study are to investigate whether these visible movements are coordinated with sounds in the production of speech and music and whether coordination depends on the intentional category of behavior rather than its physical manifestation. To illustrate, talking and singing are two different categories produced with the same physical apparatus. Talking is primarily communicative in nature, whereas singing is more expressive and evokes feelings and connection with the music. We aim to learn about the roles of multimodal coordination in communicative and expressive performances.

There are many domain-specific aspects of sound versus movement and speech versus music, but we can start to compare and relate them by first recognizing that both acoustic and motion signals vary in amplitude over time for both speech and music. Moreover, the fluctuations in amplitude are far from random—sound and movement energies are clustered in time, sometimes periodically and sometimes aperiodically but always with temporal patterning (Koelsch et al., 2013; Martin, 1972; Rohrmeier et al., 2015). Specifically, amplitude envelopes have *multiscale structure*: Acoustic energy

**CONTACT** Camila Alviar malviarguzman@ucmerced.edu Cognitive and Information Sciences, University of California, Merced, 5200 Lake Rd., Merced, CA 95343.

comes in very brief clusters (phones and notes) that are grouped together to former larger clusters (syllables and motifs), which are themselves grouped to form even larger clusters, and so on across a wide range of timescales. Multiscale structure is also found in human movements (Delignières et al., 2003, 2004; Delignières & Torre, 2009; Hausdorff et al., 1996) and can be illustrated in the movements of speech and music. For example, multiple jaw oscillations (syllables) are nested within the movement of each breath group, and multiple guitar notes and chords may be nested within each fret change.

The commonality of multiscale structure makes it an apt basis for comparing and relating multimodal coordination in speech and music. In the present study we quantified multiscale structure of sound and movements via temporal structure in the amplitude envelopes of corresponding acoustic and motion signals. Specifically, we measured nested clustering in the spectral power of amplitude envelopes, that is, the *modulation spectrum*, and compared clustering in audio and video recordings of speech and music performances. We also measured multimodal coordination more directly in the amplitude envelopes by measuring the peaks of their cross-correlation functions.

In general, we found reliable correlations in both measures of relationship between acoustic and motion signals, as expected based on previous research. However, we also observed differences in effects across different types of speech and music that provide evidence for tighter multimodal coordination during communicative performances. Next, we review the literature on multimodal coordination in speech and music, starting with perception and followed by production. We then review relevant studies of multiscale structure that taken together with the literature on multimodal coordination lead us to the present study.

### Coordination of sound and movement in speech and music

The integrated perception of sound and movement is supported by studies of both speech and music production and perception. In the case of speech production, for example, the area of the opening of the mouth at any given time has been shown to be robustly correlated with the amplitude of the speech wave 100 to 300 ms later (Chandrasekaran et al., 2009). More generally, recordings of tongue and vocal tract movements can be used to learn a causal (forward) model to produce corresponding speech sounds (Kello & Plaut, 2004).

There is also evidence that bodily movements causing little or no sound may nonetheless be coordinated with the production of speech sounds. For instance, amplitudes of movements like beat gestures of the arms have been shown to correlate with peak amplitudes in the speech signal (Pouw, Harrison et al., 2020). The temporal patterning of such amplitudes carries enough information for listeners to synchronize their own beat gestures to the speaker based on sound alone (Pouw, Paxton et al., 2020). The coupling of speech and gesture is continuously ongoing, as evidenced by a study showing that disruption in one modality quickly spreads to delay movements in the other modality (Chu & Hagoort, 2014). In fact, recent evidence shows stronger coupling of speech and gesture (i.e., more synchronization) under a delayed auditory feedback perturbation, suggesting that speech–gesture synchrony might play a role in maintaining the stability of the speech production system (Pouw & Dixon, 2019). Articulatory movements and manual gestures also reflect the broad prosodic structure present in speech, lengthening under prosodic prominence and around prosodic boundaries (Krivokapić et al., 2017) and in some cases even extending over them to indicate larger prosodic groupings (Yasinnik et al., 2004).

Studies of musical performances have similarly revealed correlations between movements performed and sounds produced. For example, the peak height of the pianists' fingers (Dalla Bella & Palmer, 2011) and the acceleration of the clarinetists' fingers (Palmer et al., 2009) both increase in response to faster tempi as way of preserving the temporal accuracy of the melody. Similarly, a study with saxophonists showed that the coordination between the tongue and the fingers was important for stability in tempo and that different articulation techniques produced sound qualities distinct enough for expert saxophonists to correctly identify the technique used to produce the melody (Hofmann & Goebl, 2014). Moreover, ancillary gestures not directly involved in producing sound have been shown to correlate with the timing, timbre, and loudness of the music, especially in moments of harmonic transition (Teixeira et al., 2018).

The link between sound and movement in speech and music can also be seen in the effects of perceived movements on perceived sounds. For instance, evidence shows that body movement affects how infants (Phillips-Silver & Trainor, 2005) and adults (Phillips-Silver & Trainor, 2007) perceive ambiguous rhythm sequences by facilitating groupings in auditory stimuli that mirror rhythms in the corresponding movements. In the case of speech, the classic McGurk effect highlights the influence of the perceived movement of the mouth on the perceived phoneme being uttered (McGurk & MacDonald, 1976). Evidence from functional magnetic resonance imaging indicates that premotor and motor cortices and the left cerebellum are involved in beat perception by simulating periodic movements that predict upcoming beats (see Gordon et al., 2018; Patel & Iversen, 2014). More generally, areas of the premotor cortex active during music production are also active during music perception, particularly in trained musicians (see Zatorre et al., 2007 for a review). An analogous effect has been found in the sensorimotor cortex during passive speech comprehension tasks (see Schomers & Pulvermüller, 2016 for a comprehensive review).

The literature reviewed thus far provides a wide range of evidence confirming and detailing the relationship between movement and sound for speech and music. However, the relationship is by no means fixed, as in gestures and other ancillary movements that may vary in their relationship to sound and aspects of sound that have complex relationships to movements not measured or perceived (e.g., changes in loudness due to subtle changes in air pressure). The variable and complex relationship between movement and sound leads us to ask whether the relationship might vary as a function of the kind of speech being spoken or music being performed. Our approach to addressing this question is based on studies, reviewed next, that found large and consistent differences in the multiscale structure of sound in different kinds of speech and musical performances, measured directly and automatically in sound recordings.

Multiscale structure refers to the organization of behavior in time or space, specifically across temporal or spatial scales of observation. Behavior is said to have multiscale structure when structured, nonrandom variability is found in behavioral signals that are windowed across a wide range of temporal or spatial resolutions. With respect to structure across temporal scales, dozens of studies have found human behavior to exhibit a particular kind of multiscale structure (Kello et al., 2010), in which temporal clustering grows in proportion with timescale. This type of multiscale structure is generally referred in the literature as *1/f scaling* (e.g., Van Orden et al., 2011) because growth of temporal structure with timescale can be expressed as an inverse relationship between spectral power and frequency. We use the more general term *hierarchical temporal structure* here to refer to a general trend for temporal structure to grow with timescale, including deviations from this trend that may be informative about underlying processes.

Most relevant to the present study, Kello et al. (2017) quantified the hierarchical temporal structure in speech and music recordings as expressed in the amplitude envelopes of the acoustic waveforms. Specifically, they extracted above-threshold peak amplitude events and quantified the degree of event clustering as a function of timescale. Short timescales measured small-scale clustering that roughly corresponded to individual phonemes and notes, whereas larger timescales measured larger-scale clustering that roughly corresponded to words and musical phrases, and so on. Allan Factor analysis (Allan, 1966) was used to quantify the pattern of clustering across timescales that corresponds to the degree and shape of hierarchical temporal structure for a given sound recording.

Kello et al. (2017) measured and compared Allan Factor functions for over 160 sound recordings, covering timescales from about 30 ms to 30 s, and showed that different genres of music (popular, classical, jazz) and different types of speech (monologue, dialogue, synthesized) could be distinguished from each other on the basis of their Allan Factor functions. Nested clustering across timescales was found to increase with musical composition, speech interaction, and prosodic emphasis. These and other communicative and expressive differences in speech and music corresponded with consistent changes in the shapes of Allan Factor functions that reflected the underlying behavioral processes (see also Ro & Kwon, 2009).

It stands to reason that hierarchical temporal structure in the sounds of speech and music may have corresponding structure in the underlying movements that produce the sounds, and indeed there is evidence in support of this conjecture. Chandrasekaran et al. (2009) found hierarchical temporal structure in spectral analyses of articulator movements during speech that suggested the nested clustering in the movements of the lips mirrors the nested clustering in the resulting speech sounds. As for gestures during speech, Pouw and Dixon (2019) used cross-wavelet analysis to show that the amplitude envelope for speech sounds is coupled with the acceleration profile of hand movements across time scales. Finally, in a recent study we used frame differencing analyses of video recordings to illustrate hierarchical temporal structure in the movement amplitudes of lecturers giving academic talks (Alviar et al., 2018).

The present study builds on prior work by analyzing structure in the sound and movement amplitude envelopes for a range of different kinds of speech and music recordings. We measure this structure in two complementary ways: one by correlating smoothed versions of the envelopes themselves and one by correlating power estimates from modulation spectra of the envelopes. We computed the acoustic amplitude envelope directly from the acoustic waveform of each recording, whereas we used video frame differencing to compute amplitude envelopes from the video signal.

Studies showing different kinds of hierarchical temporal structure for different kinds of speech and music performances (Kello et al., 2017) lead us to expect variations in multimodal, multiscale coordination depending on the category of behavior. Without more specific predictions, we start by sampling from a range of different types of speech and musical performances that are readily available on YouTube and analyzable in terms of recording conditions. We chose speech videos that varied primarily in their communicative and expressive properties, and music videos that varied primarily in their physical properties (type of instrument). We analyze whether our measures of coordination vary more as a function of functional properties in speech or physical properties in music and then present more targeted analyses based on these results.

## Methods

### Audio and video recordings

We used a convenience sample from YouTube as a source of natural data for this study. Recordings are publicly available and contain time-locked video and audio from a wide range of people and backgrounds—different ages, different culture, different gender, and different parts of the world. Convenience samples can be biased relative to the population in question. In our case, biases might arise from YouTube's search algorithm criteria or characteristics of people who have the means and interest to make and post videos, especially unproduced amateur videos. Despite these possible biases, the diversity of our sample makes it broadly generalizable and more representative of the general population than samples of college undergraduates by comparison.

The following criteria were used for selecting videos. First, we ensured that all recorded sound and movement (beyond low levels of background noise) was produced by only one performer per recording, which means the camera had to be still and shot from only one angle (hence, most recordings were amateur and not edited or produced). We did not constrain our selection of recordings based on camera angle, but most were directly facing the person being recorded, except piano recordings, which were mostly 90 degrees in profile and most captured motion at least from the torso to the head. Performances were at least 3.5 minutes so that we could analyze timescales comparable with prior studies.

Applying these selection criteria to videos available on YouTube, we initially found 80 videos, half speech and half musical performances. We deemed this number of videos sufficient for our purposes given that Kello et al. (2017) found significant differences in the shape of Allan Factor functions with a similar sample size per group. There were eight different categories of video, four for speech and four for music, with a total of 10 recordings per category. The four different categories of speech performances were chosen to sample a range of communicative styles and degrees of spontaneity: spoken word poetry, scripted acting, unscripted spontaneous monologues, and teleprompter reading. The four different types of music recordings were all classical music performances chosen to sample a range of instruments (i.e., physical means of sound production): flute, guitar, piano, or violin.

After analyzing results with these 80 recordings as presented below, we added two additional test categories, again with 10 videos each: *A cappella* singing was added to test the use of speech as an instrument, and improvisational jazz piano was added to test the use of an instrument in a more spontaneous, conversational style of performance (Kello et al., 2017). Recordings were 5.66 minutes long on average, and there were 57 male performers and 43 female performers distributed roughly evenly across the categories. Additional details about each recording are listed in Appendix 1. We did not expect to get any systematic influences of other possible common characteristics across the groups nor tested for them.

To facilitate comparisons across recordings and modality, all audio and video signals were resampled to the lowest common denominator of 30 Hz. This procedure constituted a massive down-sampling of the audio signal, which was no less than 44.1 kHz as originally recorded, into the range of typical video sample rates, which was 20 to 30 frames per second in our sample ($M = 28.45$, $SD = 2.49$). Making the sample rate a constant of 30 Hz, with a minimum recording length of 3.5 minutes, resulted in all preprocessed signals having the same range of timescales available for analysis.

## Sound and movement analyses

Figure 1 shows the basic steps of waveform analysis we took. Processing began with converting audio and video signals into amplitude envelopes using the Hilbert envelope as a standard technique for audio signals (Falk & Kello, 2017). For video signals we computed movement amplitudes using a simple frame differencing algorithm (all code available at https://github.com/camialviar/AVCoordMusicSpeech) that quantified the grayscale change (absolute differences) summed over pixels from frame to frame as a relatively instantaneous measure of overall movement amplitude (see Paxton & Dale, 2013 for a review of frame differencing techniques). Given that videos only captured movements of the performers, frame differencing combined both fine and coarse movements of the face and body. Frame differencing techniques have been found to produce comparable data to those obtained with motion tracking and computer vision techniques (Pouw et al., 2020).

### Cross-correlation peaks

One of our two measures of coordination was based on correlating the amplitude envelopes. To reduce the effects of high-frequency noise and idiosyncratic variations, the envelopes were first smoothed using a sixth-order low-pass Butterworth filter with a cutoff frequency of 0.5 Hz (red line at the top of Figure 1). We also trimmed the first ~12 seconds of data points to avoid the edge effects due to filtering. The cutoff frequency was selected empirically by inspecting the changes in the coefficient of variation of the Pearson correlation coefficients for different filter cutoffs. The 0.5-Hz cutoff was the frequency in which the decrease in the coefficient of variation stopped being linear, indicating a change in the relationship between the mean and the variability in the correlations. Cross-correlation functions were computed at lags up to ±100 data points (about 3 seconds, see top middle of Figure 1).

The cross-correlation functions can be compared directly to each other to test relative effects, but to test whether there is multimodal coordination beyond chance, we need a surrogate chance baseline. Within each category of 10 recordings, each movement amplitude series was paired with the nine sound amplitude series from each of the other recordings in the same category. We selected the highest correlation coefficients in the cross-correlation functions and averaged them across the nine surrogate pairings to obtain a unique surrogate value for each original recording. Signals of different lengths were trimmed to match the shortest one automatically as part of the cross-correlation function in R (2018).

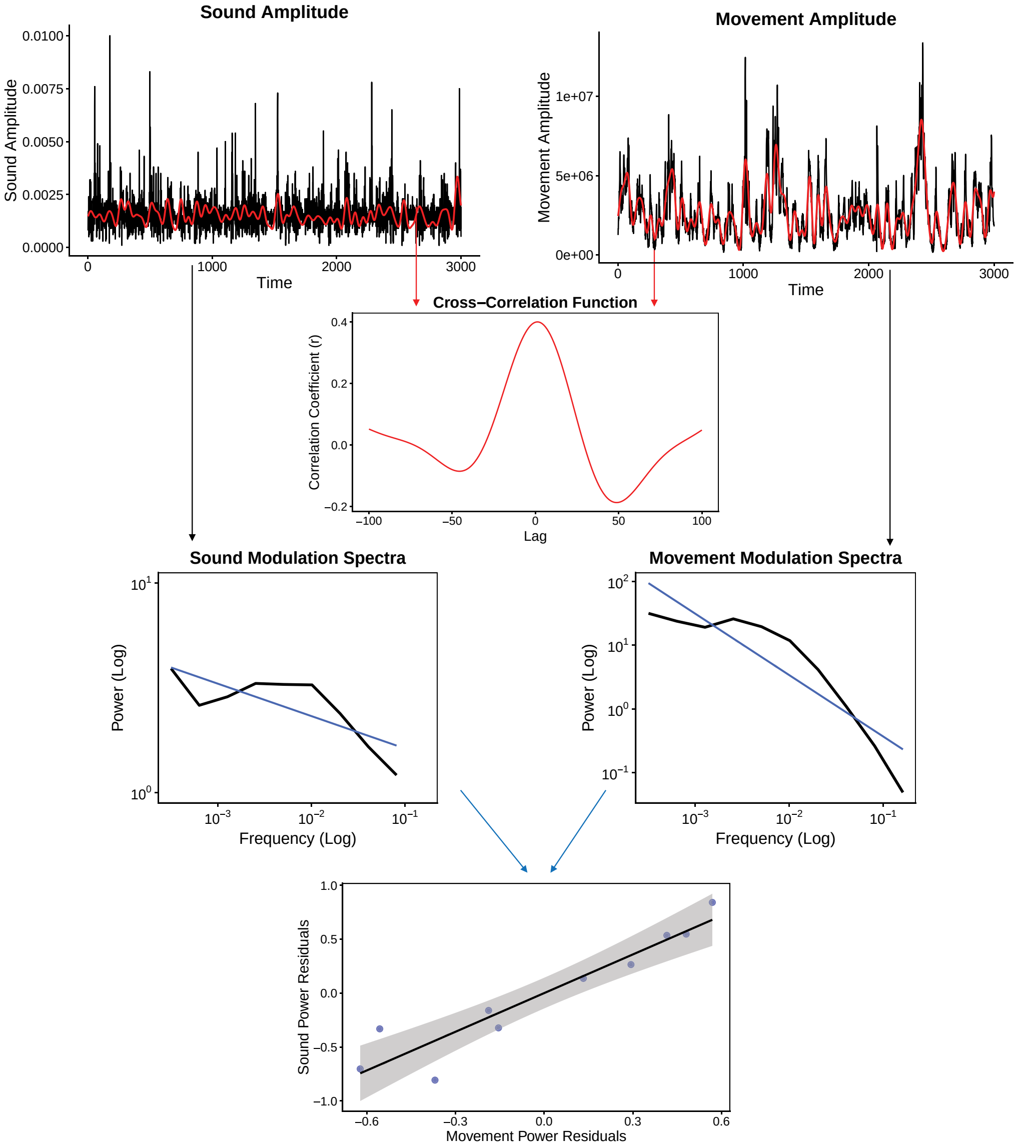

**Figure 1.** Illustration of the data analysis steps using a spontaneous speech recording as an example. Top: Downsampled sound and movement amplitude envelopes shown in black with the smoothed versions superimposed in red. Top-Middle: Cross-correlation function for the smoothed signals. Bottom-Middle: Modulation spectra of the unsmoothed amplitude signals shown in black, with the regression line for residualization superimposed in blue (*y*-axes have different ranges to better visual the similarity in spectral shapes). Bottom: Scatterplot of log-binned spectral power residuals for sound regressed onto the same for movement.

### *Modulation spectra*

Our other measure of coordination was based on correlating power estimates of the modulation spectra of unfiltered sound and movement amplitude envelopes. We applied Fourier analysis to each amplitude time series to compute spectral power estimates over the range of available frequencies (i.e., a range of timescales); hence, the multiscale nature of this analysis. The Fourier analysis decomposes a time series into sine waves at different frequencies where the amplitude of each sine wave estimates the power (amplitude squared) of each frequency in the signal. Each Fourier analysis was computed over a window of 3.5 minutes of amplitudes resampled at 30 Hz. For each recording longer than 3.5 minutes, the window was shifted forward in time, up to the recording length, and spectral power estimates were averaged over windows. Given the sample rate and set window size, the range of available frequencies for our spectral analyses was 15 Hz to 0.0048 Hz. This frequency range was divided into 10 logarithmically spaced bins, and spectral power estimates were averaged within each bin (see bottom middle of Figure 1). Logarithmic spacing meant that the lowest frequency bin contained just the one lowest frequency spectral power estimate, then the next bin averaged the next two power estimates, and so on, doubling the number of estimates averaged per bin from lowest to highest frequencies (i.e., $2^n$ spectral estimates per bin, with $n$ equal to bin number 1 to 10, low to high frequency). Logarithmic binning evens out the amount of time series data contributing to each spectral power estimate (Thornton & Gilden, 2005) and averages out idiosyncratic variance to enable comparison of spectra across modalities.

We measured the degree of spectral matching by regressing the log-binned power estimates for movement spectra onto those for sound spectra, after residualizing the effects of frequency (see regression line in bottom middle of Figure 1). Residualization was an important step because modulation spectra for speech and music are generally known to approximate a 1/f scaling relation (Voss & Clarke, 1978), as are movement time series (Alviar et al., 2018; Delignières et al., 2004; Hausdorff et al., 1996). Indeed, we replicated this well-established effect (for more details see Spectral Matching Analyses, below). Residualization was necessary to observe a correlation between sound and movement that was not driven by the general $1/f^{\alpha}$ scaling relation between power and frequency, where $0 < \alpha < 2$. A similar technique termed *complexity matching* was introduced as a measure of coordination that assumes the linear multiscale relationship in which power increases with frequency. This assumption leads one to measure complexity matching in terms of correlation between the linear coefficients of log-log regression fits (Abney et al., 2014; Coey et al., 2016; Fine et al., 2015; Marmelat & Delignières, 2012). Complexity matching was not appropriate for our purposes because we needed a measure that was sensitive to the specific bends and kinks of modulation spectra that were shown to reflect communicative and expressive aspects of speech and music (Kello et al., 2017). Therefore, we first removed the general trend of an inverse relationship between spectral power and frequency and then submitted the residuals to mixed effects models, as presented below.

Residualization did not detract from the multiscale nature of the analysis because power estimates still spanned a wide range of frequencies. Correlating residuals served as a measure of the relationship between the deviations from a 1/f trend for sounds and movements produced by the same individual. Correlated residuals reflect similarities in the hierarchical temporal structure of sound and movement amplitudes.

## Results

### Cross-correlation analyses

Figure 2 shows cross-correlation functions in each of the four main speech categories and four main music categories, for each individual recording as well as the mean of each category. At a glance the functions for speech generally have distinct peak correlations near a lag of zero, whereas functions for music have flatter profiles with more varied and less distinct peak lags. Figure 3 shows the mean peak lags and correlation coefficients for all categories. We tested for temporal coordination between sound and movement amplitudes by using peak correlation coefficients as dependent measures (Fisher Z transformed for normality) and comparing them with the mean peak coefficient for each recording's set of surrogate pairs. An analysis of variance was conducted with two independent variables, correlation type (original or surrogate) and performance type. Coefficients were stronger for original pairings ($M$ = .25) compared with surrogates ($M$ = .07), $F(1,64) = 38.136$, $p < .001$, but there were no differences between original pairings and surrogates across groups, $F(7,64) = 0.639$, $p > .7$. There was also no reliable difference between speech and music, $F(1,78) = 0.159$, $p > .6$, or between the eight different performances, $F(7,64) = 0.595$, $p > .7$. It appears that all recordings exhibited temporal

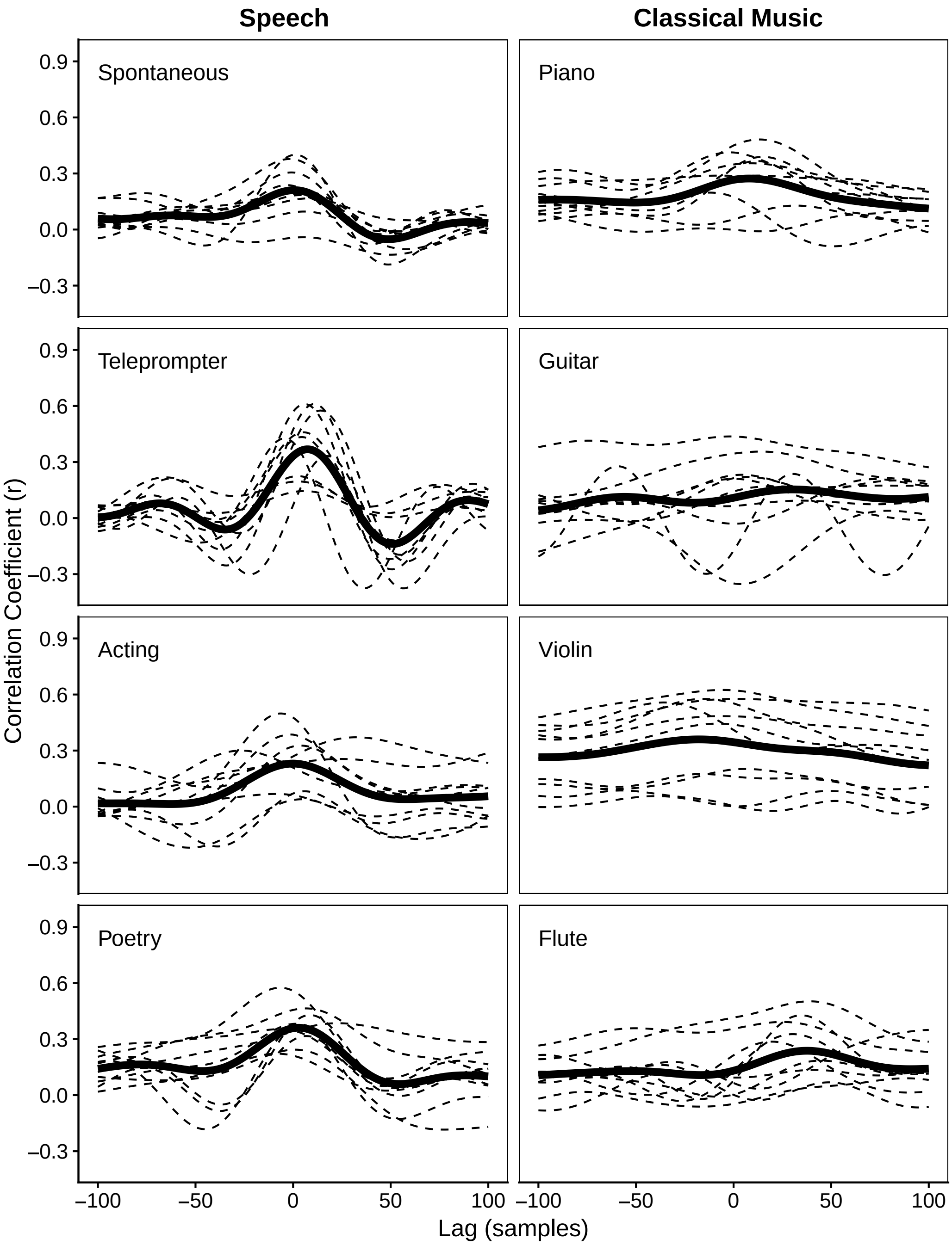

Figure 2. Cross-correlation functions for each recording category. Dotted lines show functions for individual recordings, and solid lines show category averages.

coordination above chance at some lag, but with no reliable differences in the strength of correlation between different categories of music and speech.

In contrast with correlation strengths, examination of peak lags in Figure 3 suggests a difference between categories: Lags appear relatively close to zero (*M* = 2.75, 0.09 s) for the four pure speech categories, with relatively little variation around this central tendency, indicating synchronous

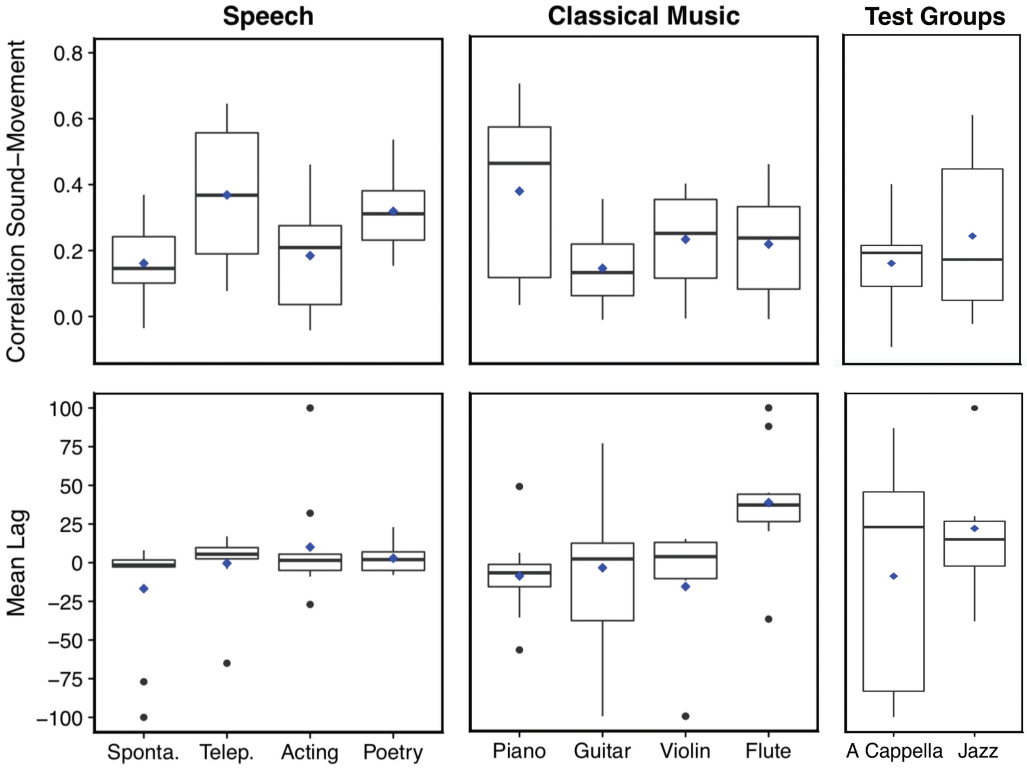

**Figure 3.** Box plots showing the distributions of peak correlation coefficients (top) and lags (bottom) for sound-movement cross-correlation functions, separated by recording category. Blue diamonds show the mean for each group. A cappella and improvisation jazz piano were analyzed separately from the main speech and music categories.

coordination of sound and movement. Lags appear different for music, in that their variances are greater both within and across subcategories. We tested this apparent difference between speech and music using two analyses. First, we conducted an F-test of equality of variances and found that the variance among peak lags for speech recordings ($S^2$ = 867) was reliably less than that for music recordings ($S^2$ = 2134), $F(39,39) = 0.406$, $p < .01$.

Second, we compared the cross-correlation function for each recording with the mean cross-correlation function of the recording's category, akin to a standardized score (e.g., a Z-score), but one that measures the consistency of an individual cross-correlation function with a mean cross-correlation function. If recordings have one or more peaks at consistent lags, then the category mean should preserve the peaks present in individual functions. By contrast, if peaks are inconsistent over recordings, then they should be averaged out in the category mean, and the mean will therefore covary less with the individual recordings. We tested these competing hypotheses in two steps: First, we correlated each individual cross-correlation function with its corresponding category mean to get a measure of consistency between each individual cross-correlation profile and the group's average profile. Second, we compared these measures of consistency across the different categories by running an analysis of variance on the Fisher Z transformed correlation coefficients obtained in the previous step. The results showed that individual recordings were more correlated with their category means for speech ($M$ = .752) compared with music ($M$ = .506), $F(1,78) = 8.779$, $p < .01$. This result supports the hypothesis that peak cross-correlation lags between sound and movement amplitudes were more consistently near zero across speech recordings compared with music recordings. Using correlation to create a kind of standardized score is unusual and may raise questions about its statistical validity, but results provide convergent evidence with the visible differences seen in the cross-correlation plots and with results from the more standard test of equality of variances.

Also, it is important to note that cross-correlation methods have been questioned because autocorrelation can interfere with the ability to determine causality of effects (Dean & Dunsmuir, 2016). In our study we aggregated cross-correlations across many time series and examined only their peaks to assess relationships, which avoids recent criticisms (for instance, see Figure 4 in Dean & Dunsmuir, 2016). Future studies may investigate causal relationships in multimodal signals using alternative methods such as transfer entropy and Granger causality.

### *Spectral matching analyses*

Figure 4 shows the mean modulation spectra for sound amplitudes and movement amplitudes in each recording category. All modulation spectra generally showed the expected inverse relationship between power and frequency in log-log coordinates (i.e., hierarchical temporal structure): Spontaneous speech $\alpha = -0.41$, Teleprompter $\alpha = -0.37$, Acting $\alpha = -0.59$, Poetry $\alpha = -0.44$, Piano $\alpha = -0.6$, Guitar $\alpha = -0.5$, Violin $\alpha = -0.54$, Flute $\alpha = -0.71$. We are unaware of prior results showing the same pattern in movement, but the ubiquity of hierarchical temporal structure in behavior provides an empirical basis for expecting it in movement as well, and indeed that is what we observed: Spontaneous speech $\alpha = -0.9$, Teleprompter $\alpha = -0.56$, Acting $\alpha = -0.85$, Poetry $\alpha = -0.84$, Piano $\alpha = -0.99$, Guitar $\alpha = -0.9$, Violin $\alpha = -0.8$, Flute $\alpha = -0.87$.

To test whether individual sound spectra were related to their corresponding movement spectra, we formulated mixed effects models to predict residual power estimates for sound (after removing the effect of frequency; see previous section) based on the same for movement. In one model we added the factor of main category (speech or music) and in another model we added subcategory as a factor (eight altogether, four speech and four music). We specified a maximal random effect structure following the recommendations of Barr et al. (2013), but the model failed to converge. We instead used a partial random effect structure including a random intercept and slope for movement residuals as a function of recording to account for multiple observations from each recording. The R notation for our model was as follows: *Residualized Sound Power* ~ *Residualized Movement Power**Group* + (1 + *Residualized Movement Power*|*Video ID*), with Group being either the two main categories or the eight subcategories. Models were run using the *lme4* package (Bates et al., 2015) and the *lmerTest* package (Kuznetsova et al., 2017) to calculate *p*-values.

We should note that when mixed effects models fail to converge, random effects related to between-subjects factors can be removed with little consequence for the Type I error rate of the model (Barr, 2013). We kept the random intercept and the random slope for movement residuals (and dropped the random slope for group) since this was the within-subjects fixed factor in our model (i.e., involved repeated measures from the same recording). The random effects structure controls for variability that comes from individual differences: It adds it to the unexplained variability of the model so the standard errors correctly reflect the unexplained variance and provide a good baseline to judge the significance of the fixed effects (Mirman, 2014). The random intercept builds into the model the assumption that different individuals will have different mean deviations from the 1/f trend in their sound signals. The random slope for movement residuals builds into the model the assumption that the strength of the relationship between sound and movement residuals will be different for different individuals.

Figure 5 shows scatterplots of sound and movement power residuals based on the mixed effects model using subcategory as a factor. The scatterplots show generally positive trend lines for individual speech recordings and more variable, inconsistent trend lines for music. A model with the main recording category (speech vs. music) as a factor, and its interaction with movement power residuals (AIC = 404.33) fits the data better than a model without main category (AIC = 415.70; $\chi^2 = 15.372$, $p < .001$). The interaction term was reliable for both speech ($B = 0.757$, $p < .001$) and music ($B = 0.255$, $p < .01$), but the relationship was significantly stronger for speech ($p < .001$). We repeated this analysis

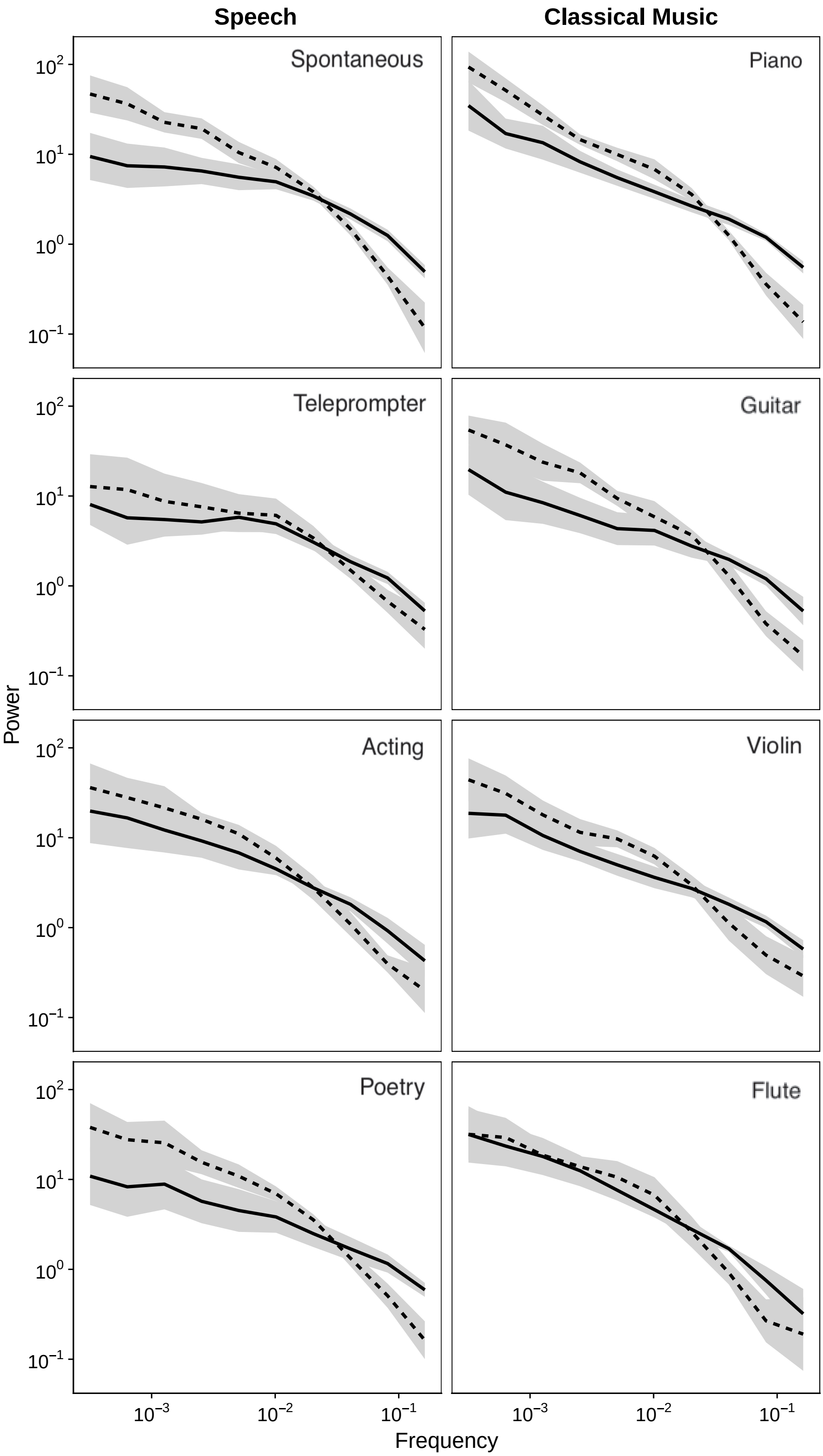

**Figure 4.** Mean modulation spectra for sound (solid line) and movement (dotted line) for each of the four main speech categories and music categories, in log-log coordinates. The shaded bands show 95% confidence intervals.

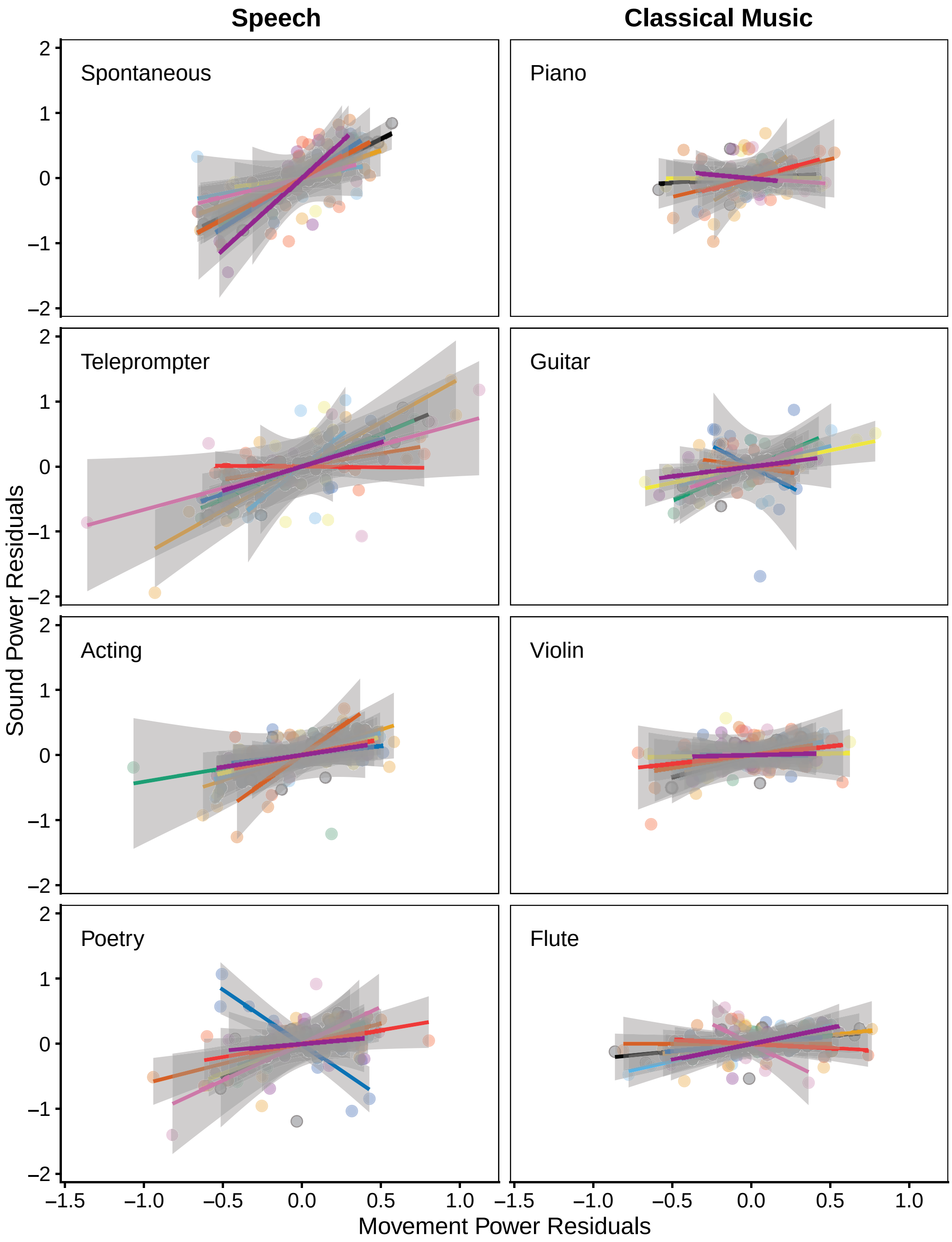

Figure 5. Movement power residuals plotted against sound power residuals based on the mixed effects model with recording subcategory as a fixed effect. Linear trend lines are shown for individual recordings in each subcategory.

with subcategory as the fixed factor instead of main category, and this time the model had a marginally worse fit with the addition of subcategory, (AIC = 418.48; $\chi^2 = 25.22$, $p < .05$). We were interested in exploring finer-grained differences among subgroups, so we planned to inspect this model regardless of its fit compared to the simpler model that did not contain subcategory. Consistent with the model including main category (speech vs. music) as a factor, all interaction terms were reliable for speech subcategories but none for classical music subcategories: Spontaneous speech ($B = 1.108$, $p < .001$), teleprompter speech ($B = 0.832$, $p < .001$), scripted acting ($B = 0.623$, $p < .001$), and poetry ($B = 0.447$, $p < .01$). These differences between the groups in speech follow a pattern that is descriptively interesting but only statistically significant at the extremes (spontaneous speech vs. poetry: $p = .009$). We should note too that the null results for the music groups are most likely the result of a lack of power to detect the smaller effects, as suggested by the significant result for music.

### *Follow-up test of category effects*

Results so far provide evidence that multimodal, multiscale coordination between movement and sound is different for speech versus musical performances. Our four main types of speech recordings showed more consistent synchronization between sound and movement amplitudes (although correlation coefficients did not differ by category), and the speech categories showed more multiscale spectral matching compared with music categories. We also found that spectral matching varied within speech categories in a pattern related to the type of communication or performance. As we mentioned above two caveats apply to this result: The differences between speech groups were only reliable at the extremes of the pattern, and the lack of differences among the music groups may be due to a lack of power.

Our category manipulation confounded speech and music with functional (communicative) differences versus physical (instrumental) differences. Given that this study is based on naturalistic observations of YouTube videos, we could not fully test these manipulations independently. However, we collected and analyzed data for two additional recording categories to provide an initial test of the effects observed thus far. One category was *a cappella* singing, which uses the vocal apparatus like speech (i.e., physical similarity) but in which the voice is used more like an instrument than a means of language communication. The other category was improvisational jazz piano, which is a musical genre that has been likened to spoken conversation (Kello et al., 2017; Sawyer, 2005).

Results for these two categories are shown in Figure 6. Singing patterned more with music than speech in that peak lags of cross-correlation functions were much more variable ($S^2 = 5194$) than those for speech ($S^2 = 867$), $F(39,9) = 0.166$, $p < .001$. Spectral matching also patterned with music in that the relationship between sound and movement spectra was weak for a cappella singing, albeit marginally reliable ($B = 0.404$, $p < .05$). Jazz also patterned with music in terms of peak lags, but the category was unique in terms of spectral matching because power residuals for sound and movement amplitudes were *negatively* correlated ($B = -0.53$, $p < .01$). On the one hand, this negative correlation is like speech because it indicates multiscale, multimodal coordination, but on the other hand, the specific mode of coordination is different from observed for speech. While further investigation is needed to understand this unique result, the two additional test categories provide further evidence that coordination between sound and movement in speech and music depends on communicative and expressive aspects of performance rather than the physical apparatus of sound production.

## Discussion

In this study we explored the coordination between sounds and movements that produce or accompany a variety of speech and musical performances. Our main objective was to compare the coordination of movement and sound for categories of speech and music as a function of functional and physical properties of the performances. In general, we found reliable relationships between movement and sound for both speech and music, as was expected from previous research. We also found reliable differences in the multimodal coordination of speech compared with music, and these differences depended on communicative and expressive aspects of performance rather than the physical mechanism of sound production.

The test group of a cappella singing showed that multiscale coordination depends on the act of speech communication rather than use of the vocal tract per se, and the test group of jazz improvisation showed that coordination depends on the musical genre rather than use of a particular instrument

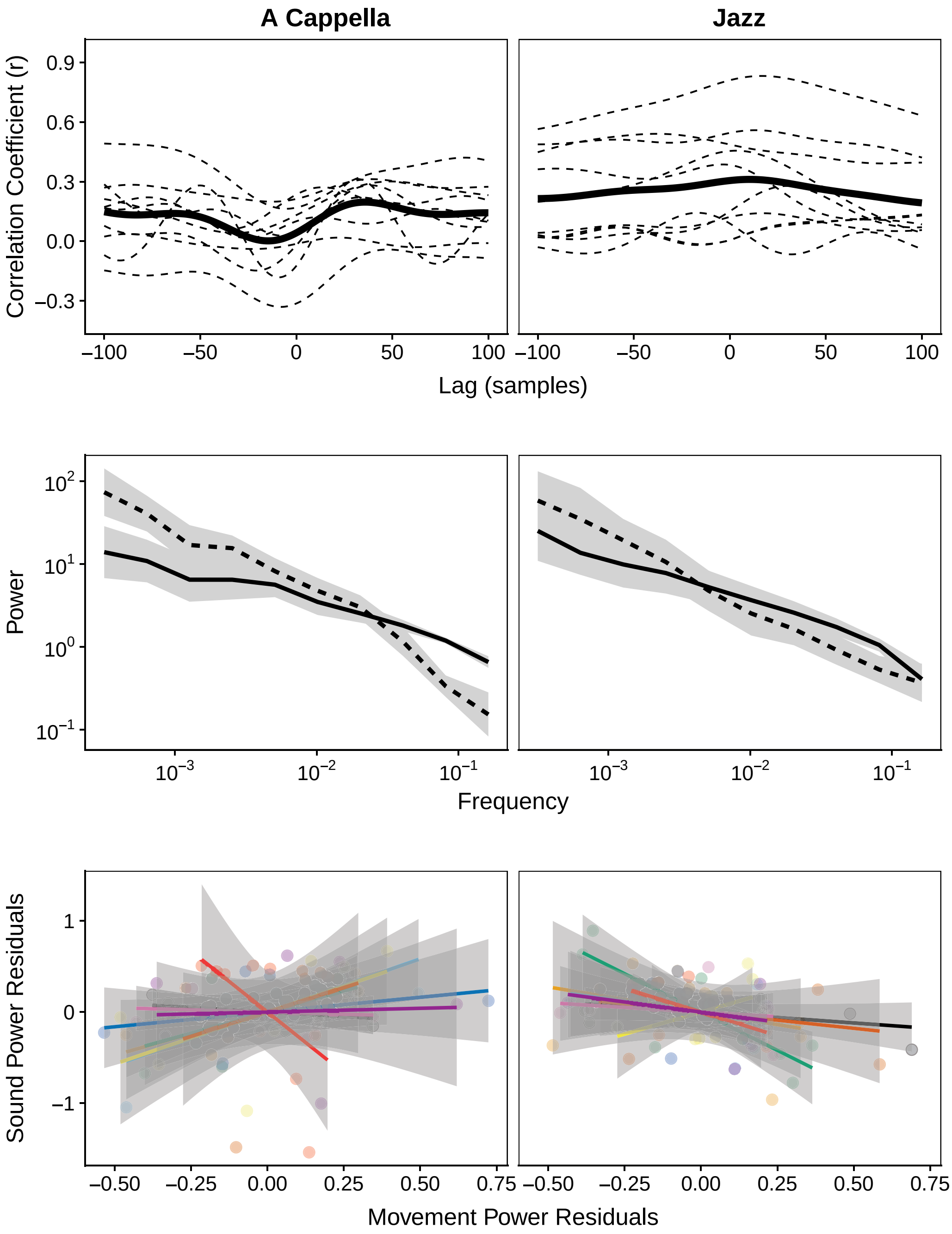

Figure 6. Top: Cross-correlation functions for A Cappella and Improvisational Jazz Piano test categories. Dotted lines show functions for individual recordings, and solid lines show category averages. Middle: Mean modulation spectra for sound (solid line) and movement (dotted line), in log-log coordinates. Shaded bands show 95% confidence intervals. Bottom: Movement power residuals plotted against sound power residuals based on the mixed effects model with recording subcategory as a fixed effect. Linear trend lines are shown for individual recordings in each subcategory.

such as the piano. Results also showed that our two measures of coordination are complementary, in that spectral matching showed greater multiscale coordination for speech compared with music, whereas the magnitude of cross-correlation showed coordination in both categories. Furthermore, the lags of peaks in cross-correlations highlighted the relative synchrony of sound-movement coordination in speech but not music, and both measures were necessary to show the unique pattern of coordination of jazz compared with speech and classical music.

One might wonder if the differences between speech and music may be attributable to the way movements are captured differently in audio or video signals. However, the nature of recordings was largely the same across subcategories of speech and music, yet subcategory differences were found herein and also in Kello et al. (2017) for sound recordings using Allan Factor analysis. Also, across categories there was no appreciable difference between videos of speech performances and singing, and yet their coordination patterns were distinct. In both categories performers used their vocal tracts and were filmed from the same basic angle and distance.

We find a functional explanation to be more consistent with the results. In particular, speech performances served the purpose of language communication, whereas musical performances did not. McNeill (1985) showed that gestures mirror and complement the meanings and discursive organization of speech communication, and listeners expect gesture to carry communicative information. Speakers gesture more when talking to an interlocutor (Bavelas et al., 2008), they adapt their gestures to the needs of their audiences (Galati & Brennan, 2014; Özyürek, 2002), and they use gestures purposefully to disambiguate ambiguous sentences for the listener (Holler & Beattie, 2003). Speakers' gestures also carry relevant information about the referents in the narrative aiding discourse comprehension (Debreslioska et al., 2013) and help the listener gauge the speaker's confidence on the information being conveyed (Roseano et al., 2016). Listeners are slower and are less accurate to match meanings when presented with incongruent speech–gesture pairs (e.g., hearing "chop" and seeing a "twist" gesture) even when the task requires attention to just one of the modalities (Kelly et al., 2010), and listeners show differences in event-related potentials (the N400 associated with semantic irregularities in sentences) for semantically congruent versus incongruent speech–gesture pairs (Habets et al., 2011). Listeners also take advantage of congruent gestures to facilitate processing of ambiguous speech, as evidenced in the reduction of the N400 for ambiguous sentences preceded by a disambiguating gesture (Holle & Gunter, 2007). This prior research suggests an intimate temporal relationship between speech and co-occurring signals, such as gestures. The relationship is profoundly functional in the sense that the success of communication depends on this dynamic relationship. Thus, evidence for greater multiscale coordination in speech may stem from the demands of speech communication.

Our functional interpretation leads to a general prediction that may be tested in future studies. First, an increase on communicative demands should increase the strength of the multiscale, multimodal coordination in speech. For instance, communication in more constraining or noisy conditions should lead to stronger spectral matching and greater synchronicity between sound and movement (cf., Boker et al., 2002; Paxton & Dale, 2017). Similarly, the communication of more elaborate, complex messages should lead to the same pattern of effects. The same logic may be applied to jazz improvisation, which has been likened to a conversational interaction (Kello et al., 2017; Sawyer, 2005). Improvisation among two or more musicians should yield greater multiscale, multimodal coordination due to communication among the musicians compared with solo performances. Musical gestures have been shown to play a communicative role in a string quartet leading to improvement of their performance (Hospelhorn & Radinsky, 2017). And in the context of jazz, a previous study suggests that movement coordination between two jazz players engaged in improvisation follows complex and interesting coordination patterns of their heads and their arms at different time scales (Walton et al., 2015). The negative relationship of the spectral power in jazz suggests that decoupling movements and sounds might be a strategy to perform creatively in jazz and perhaps also in other genres. Doing similar analyses for improvisation in other musical genres and nonimprovisational jazz will help get a clearer picture of multimodal coordination across musical performances.

Our functional interpretation is also supported by differences in magnitudes of spectral matching among the subcategories of speech performances. Beta coefficients patterned with the degree of message information carried by the movements. Movements during spontaneous speech appeared to convey the most information about the content of the message, followed by teleprompter speeches. By contrast, movements during monologue acting and spoken poetry appeared to be more associated with emotional, stylistic aspects of performance. Spoken poetry, for example, emphasized beat gestures and rhythmic movements that reflect the prosody of the poem more than its semantic content. These observations are admittedly subjective and require more systematic experimentation to test them.

Another possible factor that might affect multiscale, multimodal coordination is the degree of spontaneity in a performance. Previous studies have indicated that patterns of prosodic variation and gesture production for spoken poetry, for example, are more tightly scripted than they are for spontaneous conversation (see Novak, 2011; Barney, 1999 for an analysis of the delivery of spoken word; Henning, 1955 for an analysis of speech delivery). By contrast, spontaneous speech and musical improvisation, for instance, may result in stronger synergies to reduce the additional degrees of freedom and ease coordination (see Kelso, 2009). Experiments that evoke different combinations of spontaneity and communication could help delineate and tease apart their influences on coordination. For instance, academic presentations may be more scripted compared with spontaneous retellings of cartoons (see Galati & Brennan, 2014).

In summary, our study shows how naturalistic recordings can provide a wealth of information about speech, music, and other human behaviors as they occur “in the wild.” We analyzed YouTube recordings and found evidence that the coordination of sound and movement is more consistent and synchronous in speech than in music, and this conclusion seems to depend on functional aspects of the performances and not the physical apparatus per se. The relevant functional factors may be related to communicative constraints, degree of spontaneity, and the general properties of language communication as a multiscale, multimodal performance (Dale & Kello, 2018). Future studies may test these hypotheses with more experimental control to minimize possible differences in recording conditions that might exist despite our careful selection criteria and controls.

## Disclosure statement

The authors declare no conflict of interests.

## Data availability statement

The datasets and scripts used in the data analyses are available in the GitHub repository https://github.com/camialviar/AVCoordMusicSpeech.

## References

Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2014). Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General*, *143*(6), 2304. https://doi.org/10.1037/xge0000021

Allan, D. W. (1966). Statistics of atomic frequency standards. *Proceedings of the IEEE*, *54*(2), 221–230. https://doi.org/10.1109/PROC.1966.4634

Alviar, C., Dale, R., & Kello, C. (2018, July). The fractal structure of extended communicative performance. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1292–1297). Austin, TX: Cognitive Science Society.

Barney, T. (1999). Readers as text processors and performers: A new formula for poetic intonation. *Discourse Processes*, *28*(2), 155–167. https://doi.org/10.1080/01638539909545078

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328. https://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495–520. https://doi.org/10.1016/j.jml.2007.02.004

Bernstein, N. A. (1967). *The coordination and regulation of movements*. Pergamon Press.

Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, *7*(3), 338. https://doi.org/10.1037/1082-989X.7.3.338

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Chu, M., & Hagoort, P. (2014). Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General*, *143*(4), 1726–1741. https://doi.org/10.1037/a0036281

Coey, C. A., Washburn, A., Hassebrock, J., & Richardson, M. J. (2016). Complexity matching effects in bimanual and interpersonal syncopated finger tapping. *Neuroscience Letters*, *616*, 204–210. https://doi.org/10.1016/j.neulet.2016.01.066

Dale, R., & Kello, C. T. (2018). "How do humans make sense?" Multiscale dynamics and emergent meaning. *New Ideas in Psychology*, *50*, 61–72. https://doi.org/10.1016/j.newideapsych.2017.09.002

Dalla Bella, S., & Palmer, C. (2011). Rate effects on timing, key velocity, and finger kinematics in piano performance. *PloS One*, *6*(6), e20518. https://doi.org/10.1371/.journal.pone.0020518

Dean, R. T., & Dunsmuir, W. T. (2016). Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior Research Methods*, *48*(2), 783–802. https://doi.org/10.3758/s13428-015-0611-2

Debreslioska, S., Özyürek, A., Gullberg, M., & Perniss, P. (2013). Gestural viewpoint signals referent accessibility. *Discourse Processes*, *50*(7), 431–456. https://doi.org/10.1080/0163853X.2013.824286

Delignières, D., Deschamps, T., Legros, A., & Caillou, N. (2003). A methodological note on nonlinear time series analysis: Is the open-and closed-loop model of Collins and De Luca (1993) a statistical artifact? *Journal of Motor Behavior*, *35*(1), 86–96. https://doi.org/10.1080/00222890309602124

Delignières, D., Lemoine, L., & Torre, K. (2004). Time intervals production in tapping and oscillatory motion. *Human Movement Science*, *23*(2), 87–103. https://doi.org/10.1016/j.humov.2004.07.001

Delignières, D., & Torre, K. (2009). Fractal dynamics of human gait: A reassessment of the 1996 data of Hausdorff et al. *Journal of Applied Physiology*, *106*(4), 1272–1279. https://doi.org/10.1152/japplphysiol.90757.2008

Falk, S., & Kello, C. T. (2017). Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition*, *163*, 80–86. https://doi.org/10.1016/j.cognition.2017.02.017

Fine, J. M., Likens, A. D., Amazeen, E. L., & Amazeen, P. G. (2015). Emergent complexity matching in interpersonal coordination: Local dynamics and global variability. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 723. https://doi.org/10.1037/xhp0000046

Galati, A., & Brennan, S. E. (2014). Speakers adapt gestures to addressees' knowledge: Implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, *29*(4), 435–451. https://doi.org/10.1080/01690965.2013.796397

Gordon, C. L., Cobb, P. R., & Balasubramaniam, R. (2018). Recruitment of the motor system during music listening: An ALE meta-analysis of fMRI data. *PloS One*, *13*(11), e0207213. https://doi.org/10.1371/journal.pone.0207213

Habets, B., Kita, S., Shao, Z., Özyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*(8), 1845–1854. https://doi.org/10.1162/jocn.2010.21462

Hausdorff, J. M., Purdon, P. L., Peng, C. K., Ladin, Z. V. I., Wei, J. Y., & Goldberger, A. L. (1996). Fractal dynamics of human gait: Stability of long-range correlations in stride interval fluctuations. *Journal of Applied Physiology*, *80*(5), 1448–1457. https://doi.org/10.1152/jappl.1996.80.5.1448

Henning, J. H. (1955). How to deliver a speech. *Communication Quarterly*, *3*(1), 3–21. https://doi.org/10.1080/01463375509384871

Hofmann, A., & Goebl, W. (2014). Hierarchical organization in the temporal structure of infant-direct speech and song. Production and perception of legato, portato, and staccato articulation in saxophone playing. *Frontiers in Psychology*, *5*, 690. https://doi.org/10.3389/fpsyg.2014.00690

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*(7), 1175–1192. https://doi.org/10.1162/jocn.2007.19.7.1175

Holler, J., & Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener? *Gesture*, *3*(2), 127–154. https://doi.org/10.1075/gest.3.2.02hol

Hospelhorn, E., & Radinsky, J. (2017). Method for analyzing gestural communication in musical groups. *Discourse Processes*, *54*(7), 504–523. https://doi.org/10.1080/0163853X.2015.1137183

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*(5), 223–232. https://doi.org/10.1016/j.tics.2010.02.005

Kello, C. T., Dalla Bella, S., Médé, B., & Balasubramaniam, R. (2017). Hierarchical temporal structure in music, speech and animal vocalizations: Jazz is like a conversation, humpbacks sing like hermit thrushes. *Journal of the Royal Society Interface*, *14*(135), 20170231. https://doi.org/10.1098/rsif.2017.0231

Kello, C. T., & Plaut, D. C. (2004). A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *The Journal of the Acoustical Society of America*, *116*(4), 2354–2364. https://doi.org/10.1121/1.1715112

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267. https://doi.org/10.1177/0956797609357327

Kelso, J. S. (2009). Synergies: Atoms of brain and behavior. In D. Sternad (Ed.) *Progress in Motor Control. Advances in Experimental Medicine and Biology* (Vol. 629, pp. 83–91). Boston, MA: Springer.

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences*, *110*(38), 15443–15448. https://doi.org/10.1073/pnas.1300272110

Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1), 3. https://doi.org/10.5334/labphon.75

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

Marmelat, V., & Delignières, D. (2012). Strong anticipation: Complexity matching in interpersonal coordination. *Experimental Brain Research*, *222*(1–2), 137–148. https://doi.org/10.1007/s00221-012-3202-9

Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review*, *79*(6), 487–509. https://doi.org/10.1037/h0033467

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746. https://doi.org/10.1038/264746a0

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, *92*(3), 350. https://doi.org/10.1037/0033-295X.92.3.350

Mirman, D. (2014). *Growth curve analysis and visualization using R*. CRC Press.

Novak, J. (2011). *Live poetry: An integrated approach to poetry in performance* (Vol. 153). Rodopi.

Özyürek, A. (2002). Do speakers design their co-speech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language*, *46*(4), 688–704. https://doi.org/10.1006/jmla.2001.2826

Palmer, C., Koopmans, E., Loehr, J. D., & Carter, C. (2009). Movement-related feedback and temporal accuracy in clarinet performance. *Music Perception: An Interdisciplinary Journal*, *26*(5), 439–449. https://doi.org/10.1525/mp.2009.26.5.439

Patel, A. D., & Iversen, J. R. (2014). The evolutionary neuroscience of musical beat perception: The Action Simulation for Auditory Prediction (ASAP) hypothesis. *Frontiers in Systems Neuroscience*, *8*, 57. https://doi.org/10.3389/fnsys.2014.00057

Paxton, A., & Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior Research Methods*, *45*(2), 329–343. https://doi.org/10.3758/s13428-012-0249-2

Paxton, A., & Dale, R. (2017). Interpersonal movement synchrony responds to high-and low-level conversational constraints. *Frontiers in Psychology*, *8*, 1135. https://doi.org/10.3389/fpsyg.2017.01135

Phillips-Silver, J., & Trainor, L. J. (2005). Feeling the beat: Movement influences infant rhythm perception. *Science*, *308*(5727), 1430. https://doi.org/10.1126/science.1110922

Phillips-Silver, J., & Trainor, L. J. (2007). Hearing what the body feels: Auditory encoding of rhythmic movement. *Cognition*, *105*(3), 533–546. https://doi.org/10.1016/j.cognition.2006.11.006

Pouw, W., & Dixon, J. A. (2019). Entrainment and modulation of gesture–speech synchrony under delayed auditory feedback. *Cognitive Science*, *43*(3), e12721. https://doi.org/10.1111/cogs.12721

Pouw, W., Harrison, S. J., & Dixon, J. A. (2020). Gesture-speech physics: The biomechanical basis of gesture-speech synchrony. *Journal of Experimental Psychology: General*, *149*(2), 391–404. https://doi.org/10.1037/xge0000646

Pouw, W., Paxton, A., Harrison, S. J., & Dixon, J. A. (2020). Acoustic specification of upper limb movement in voicing. *PNAS*, *117*(21), 11364–11367. https://doi.org/10.1073/pnas.2004163117

Pouw, W., Trujillo, J., & Dixon, J. A. (2020). The quantification of gesture-speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*, *52*, 723–740. https://doi.org/10.3758/s13428-019-01271-9

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ro, W., & Kwon, Y. (2009). 1/f noise analysis of songs in various genre of music. *Chaos, Solitons, and Fractals*, *42*(4), 2305–2311. https://doi.org/10.1016/j.chaos.2009.03.129

Rohrmeier, M., Zuidema, W., Wiggins, G. A., & Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664), 20140097. https://doi.org/10.1098/rstb.2014.0097

Roseano, P., González, M., Borràs-Comes, J., & Prieto, P. (2016). Communicating epistemic stance: How speech and gesture patterns reflect epistemicity and evidentiality. *Discourse Processes*, *53*(3), 135–174. https://doi.org/10.1080/0163853X.2014.969137

Sawyer, R. K. (2005). Music and conversation. *Musical Communication*, *45*, 60. http://doi.org/10.1093/acprof:oso/9780198529361.001.0001

Schomers, M. R., & Pulvermüller, F. (2016). Is the sensorimotor cortex relevant for speech perception and understanding? An integrative review. *Frontiers in Human Neuroscience*, *10*, 435. https://doi.org/10.3389/fnhum.2016.00435

Teixeira, E. C., Loureiro, M. A., & Yehia, H. C. (2018). Expressiveness in music from a multimodal perspective. *Music Perception: An Interdisciplinary Journal*, *36*(2), 201–216. https://doi.org/10.1525/MP.2018.36.2.201

Thornton, T. L., & Gilden, D. L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin & Review*, *12*(3), 409–441. https://doi.org/10.3758/BF03193785

Van Orden, G. C., Kloos, H., & Wallot, S. (2011). Living in the pink: Intentionality, wellbeing, and complexity. In C. Hooker (Ed.), *Philosophy of complex systems* (pp. 629–672). Elsevier: North Holland.

Voss, R. F., & Clarke, J. (1978). "1/f noise" in music: Music from 1/f noise. *The Journal of the Acoustical Society of America*, *63*(1), 258–263. https://doi.org/10.1121/1.381721

Walton, A. E., Richardson, M. J., Langland-Hassan, P., & Chemero, A. (2015). Improvisation and the self-organization of multiple musical bodies. *Frontiers in Psychology*, *6*, 313. https://doi.org/10.3389/fpsyg.2015.00313

Yasinnik, Y., Renwick, M., & Shattuck-Hufnagel, S. (2004). The timing of speech-accompanying gestures with respect to prosody. In *Proceedings of the international conference: From sound to sense* (pp. 97–102). Cambridge, MA: MIT.

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory–motor interactions in music perception and production. *Nature Reviews Neuroscience*, *8*(7), 547. https://doi.org/10.1038/nrn2152

**Appendix 1.** Video Information and Links

| Group | Short Link | Content Description | Duration |
|---|---|---|---|
| Spontaneous speech | http://y2u.be/rq6pnckNFIU | Male giving tips about being natural on camera | 9:23 |
| | http://y2u.be/jD5BauEdw3A | Male giving tips about being natural on camera | 6:21 |
| | http://y2u.be/tsUKwhxd0jY | Male giving tips about being natural on camera | 4:51 |
| | http://y2u.be/zNNm_szsK2g | Male sharing his thoughts on race | 5:20 |
| | http://y2u.be/4NriDTxseog | Female showcasing a variety of accents | 6:09 |
| | http://y2u.be/xhaxeuhzoRg | Male giving tips on dating | 5:10 |
| | http://y2u.be/RYrGDo4uIQU | Male sharing his feelings and thoughts | 5:52 |
| | http://y2u.be/9jpdFMi8_Gk | Female telling people about her day | 5:31 |
| | http://y2u.be/Tb2XGL-Jpqc | Female giving tips on weight loss | 5:23 |
| | http://y2u.be/n7Tth9bcgSo | Male giving his thoughts on a product | 4:45 |
| Teleprompter reading | http://y2u.be/XOLqVIkUqPY | Female responding to the State of the Union in 2015 | 6:55 |
| | http://y2u.be/wdwVc59L0zo | Male responding to the State of the Union in 2011 | 6:36 |
| | http://y2u.be/rvbRfbI6Fhk | Male responding to the State of the Union in 2013 | 9:54 |
| | http://y2u.be/NLmZbBh83-I | Male responding to the State of the Union in 2013 | 9:05 |
| | http://y2u.be/ynVg2lLF3fU | Male responding to the State of the Union in 2019 | 4:36 |
| | http://y2u.be/Qkaay1P7Ar8 | Female responding to the State of the Union in 2014 | 8:31 |
| | http://y2u.be/m13KmA-zsoU | Female responding to the State of the Union in 2016 | 8:30 |
| | http://y2u.be/QFK8aTpYAmg | Male responding to the Address to Congress in 2009 | 8:22 |
| | http://y2u.be/sTjI0FdkyTo | Male responding to the State of the Union in 2019 | 5:44 |
| | http://y2u.be/j9fts21s3Ao | Male responding to the State of the Union in 2015 | 7:06 |
| Acting monologue | http://y2u.be/IndJdgbQQPg | Male performing Oedipus | 4:51 |
| | http://y2u.be/lc5M3BVDO60 | Female performing a dramatic monologue | 3:27 |
| | http://y2u.be/uLwWk6WppVg | Female performing a dramatic monologue | 3:32 |
| | http://y2u.be/W219EQiRwHY | Male performing a dramatic monologue | 3:36 |
| | http://y2u.be/PWpGlGpDi0o | Male performing a dramatic monologue | 4:57 |
| | http://y2u.be/MbJmc9XgpxU | Female performing a dramatic monologue | 5:56 |
| | http://y2u.be/LhzxdJPnUc4 | Female performing a monologue | 3:52 |
| | http://y2u.be/Gpfkwaz2qzA | Male performing a comedic monologue | 4:48 |
| | http://y2u.be/08XlIZLE87Y | Male performing a dramatic monologue | 5:16 |
| | http://y2u.be/XYdN17xcJic | Male performing a dramatic monologue | 4:14 |

(*Continued*)

**Appendix 1.** (Continued).

| | | | |
|---|---|---|---|
| Poetry | http://y2u.be/qnaNA_cSy4s | Male performing Christian spoken word | 5:31 |
| | http://y2u.be/wUmO5h53MjA | Female performing spoken word about forgiveness | 4:52 |
| | http://y2u.be/6NU55c0irY0 | Female performing spoken word about black love | 5:19 |
| | http://y2u.be/Re9P0CzNzqk | Female performing Christian spoken word | 5:12 |
| | http://y2u.be/xDMSWrMtgm8 | Female performing spoken word about love | 5:01 |
| | http://y2u.be/c0jwZpmEILg | Female performing spoken word about sexual abuse | 5:18 |
| | http://y2u.be/60bmnZh5M80 | Male performing spoken word about love | 4:37 |
| | http://y2u.be/PC41oQWJKDw | Female performing spoken word about weed | 4:01 |
| | http://y2u.be/v-qhk5NGFlM | Female performing spoken words about love | 4:04 |
| | http://y2u.be/5Z14d5cw688 | Male performing spoken word about love | 3:47 |
| Classical piano | http://y2u.be/FpdiVtSIl_c | Female performing Liszt' Hungarian Rhapsody no. 14 | 5:53 |
| | http://y2u.be/4J52RSo0m9k | Female performing Debussy's Reverie | 5:32 |
| | http://y2u.be/mVxQOy0US9o | Male performing Mendelssohn's Rondo Capriccioso | 6:59 |
| | http://y2u.be/rsAm3pCPeRU | Male performing Impromptu no. 1 | 9:58 |
| | http://y2u.be/Fw-CS4Oe8F4 | Female performing Mozart's Sonata XIII | 5:07 |
| | http://y2u.be/qNzFpN1h1bo | Male performing Beethoven's Sonata | 7:31 |
| | http://y2u.be/hzRWvPMzyCU | Male performing Joplin's Solace | 6:32 |
| | http://y2u.be/6LZBy8lpICA | Female performing Grieg and Masquerade's Nocturne | 4:26 |
| | Link unavailable. Video available upon request | Female performing a piano piece | 5:35 |
| | http://y2u.be/z57BQZQBXi0 | Female performing Debussy's L'Isle Joyeuse | 5:49 |
| Classical guitar | http://y2u.be/EaLuZ0mYMKc | Male performing Bach's Cello Suite VI | 4:06 |
| | http://y2u.be/Xnpjm5ybOml | Male performing Hallelujah | 5:31 |
| | http://y2u.be/olW6-jhSgMg | Female performing Bach's Sonata II | 7:12 |
| | http://y2u.be/50AlHGCbevl | Female performing Bach | 4:20 |
| | http://y2u.be/q932l_1HRIQ | Male performing Satie's Gymnopedie 1 | 4:06 |
| | http://y2u.be/mu3Pu_sav5I | Female performing Dowland's Fantasia | 3:39 |
| | http://y2u.be/FVkXmaRanmM | Male performing Tansman's Variations | 8:31 |
| | http://y2u.be/N-XORX1Nr8M | Male performing Bach's Preludio | 4:46 |
| | http://y2u.be/mf2ihqYPwJU | Male performing Rodrigo's Invocacion y danza | 5:25 |
| | http://y2u.be/m4wSFFcYKnw | Male performing Aguado's Rondo Brillante 2 | 5:39 |
| Classical violin | http://y2u.be/NXPEClCFoTY | Female performing Bach's Sonata II | 9:09 |
| | http://y2u.be/X-F5jlG1Tvo | Female performing Bach's Sonata I | 5:26 |
| | http://y2u.be/vCS0pHKYcrl | Female performing Bach's Sonata I | 5:27 |
| | http://y2u.be/5tl583Yf7nQ | Male performing Mozart's Concerto no. 3 | 6:45 |
| | http://y2u.be/sGY-byNy7DA | Female performing various orchestral excerpts | 5:14 |
| | http://y2u.be/xbkPX2vTqaE | Female performing various orchestral excerpts | 6:52 |
| | http://y2u.be/zRi1NwshMbs | Male performing Mozart's Concerto | 7:11 |
| | http://y2u.be/rvPpsAbkNYw | Male performing for an audition | 6:13 |
| | http://y2u.be/OwH2BkqCTxo | Male performing for an audition | 3:45 |
| | http://y2u.be/srek6IjDuWk | Male performing Bach's Partita no. 3 | 4:13 |
| Classical flute | http://y2u.be/slJibwlhpgI | Male performing unknown melody | 3:28 |
| | http://y2u.be/Wj8it_ad0H0 | Male performing Mozart's Concerto | 7:34 |
| | http://y2u.be/0hDTi3OYF38 | Female performing Bach's Sonata | 10:00 |
| | http://y2u.be/wvBp4WTbozg | Female performing Ibert's Piece for flute solo | 5:22 |
| | http://y2u.be/J0KLwTqa4UM | Female performing Honegger's La danse de la chèvre | 3:44 |
| | http://y2u.be/S7wdOAkjgNw | Female performing Yeon Lee's Sanjo | 4:07 |
| | http://y2u.be/GxKHNYlZeRY | Male performing Celtic themes | 6:54 |
| | http://y2u.be/vk_4oz3W_xg | Male performing Hotteterre's Echoes | 5:02 |
| | http://y2u.be/PNQ8UQ2Pmbk | Female performing Yokoyama's Siren Sorrento | 3:57 |
| | http://y2u.be/kDdt-Sb0Rzg | Female performing Karg Elert's Capriccio | 3:30 |

(*Continued*)

Appendix 1. (Continued).

| | | | |
|---|---|---|---|
| A cappella singing | http://y2u.be/1UAXjz-Hioo | Female singing Pink's Perfect | 3:34 |
| | http://y2u.be/cGScZzMJjuQ | Male singing Sam Smith's Lay me down | 4:39 |
| | http://y2u.be/CwS2FtnyZBE | Female singing Adele's Hello | 4:44 |
| | http://y2u.be/JSn5xF24_U0 | Female singing Adele's Someone like you | 4:29 |
| | http://y2u.be/tpYtA_m7Ga8 | Male singing Houston's I will always love you | 4:01 |
| | http://y2u.be/fC7xhBnU_dc | Female singing Kiara's Gold | 4:44 |
| | http://y2u.be/LBH5yogi5mo | Female singing Edges' Lying There | 4:07 |
| | http://y2u.be/TSxMccDEfpM | Female singing a Nirvana medley | 5:54 |
| | http://y2u.be/vt-zb6tffp8 | Female singing Spice Girls' Mama | 4:27 |
| | http://y2u.be/Ps6q0J7hfpc | Female singing Houston's I will always love you | 4:19 |
| Improvisational Jazz piano | http://y2u.be/6EfwdGdydFQ | Male improvising jazz | 8:18 |
| | http://y2u.be/z3msPpkoVp4 | Male improvising jazz | 4:20 |
| | http://y2u.be/XZfXkuOhhLg | Male improvising jazz | 5:33 |
| | http://y2u.be/a-gTw5kExQM | Male improvising jazz | 5:08 |
| | http://y2u.be/x7_3i7C3ZE0 | Male improvising jazz | 8:23 |
| | http://y2u.be/SKKYCHHbWI8 | Male improvising jazz | 9:25 |
| | http://y2u.be/plxkLqH2u0I | Male improvising jazz | 8:06 |
| | http://y2u.be/Xbl_n0paCd4 | Male improvising jazz | 3:33 |
| | http://y2u.be/Ey6bx4kwk2Y | Male improvising jazz | 5:11 |
| | http://y2u.be/XZfXkuOhhLg | Male improvising jazz | 5:33 |