# Using Novel Dynamic Modeling Technique to Explore Synchrony of Facial Expressions and Speech in Dyadic Conversations

M. Joseph Meyer

Shelby Township, MI

Master's of Art, Psychology
Bachelor's of Science, Mathematics
Bachelor's of Art, German Modified

A Dissertation presented to the Graduate Faculty of
the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Psychology

University of Virginia
April, 2018

Committee Members:
Chair: Steve Boker (Chair)
Cynthia Tong
Karen Schmidt
Luke Dahl (Department of Music)

# Abstract

Communication and conversation are important human behaviors and have been modeled before. However, very little research has encompassed the full dynamics of and the relationship between facial expressions and speech between people measured intensively across time, including the processes of turn-taking, delays between participants, and synchrony. Furthermore, these dynamics may be best represented in a multidimensional framework, which is not typically used in this research. In this work, I introduce the novel method of windowed canonical correlation analyses in order to be able to analyze the relationships between two sets of intense longitudinal multidimensional data, such as data from two individuals in a conversation. I then apply this method and windowed cross-correlations to point coordinates from motion tracked faces and amplitudes across frequencies from transformed speech in order to explore whether synchrony and turn-taking can be found between facial expressions and speech in unscripted dyadic conversations. After performing these analyses, the speech component of conversations was found to drive the correlations between speech and faces and play a large part in the dynamics of the conversations. Limited evidence of synchrony and some evidence towards turn-taking were also found between individuals.

# List of Figures

# Table of Contents

# 1. Literature Review

## 1.1 Introduction

Communication is a necessary aspect of everyday life. Whether we're on the phone, on social media, or sitting down with someone in a coffee shop, we use different forms of communication in our lives. In face-to-face conversation, both nonverbal and verbal processes are used as part of communication. Nonverbal processes include facial expressions and other head and body movement, and verbal processes include acoustic and linguistic patterns. One interesting relationship between nonverbal and verbal behaviors in conversations is between facial expressions and nonlinguistic components of voices. This is likely due to the direct association between mouth and other facial movement and auditory speech patterns, both on a physical level and on a psychological level.

As participants of conversations talk to each other, they may mimic, sometimes subconsciously, their partner's movement and speech. This mimicry is likely to maximize their understanding of what they learn from their partner, as well as to maximize the communication of their own feelings and thoughts. Two participants moving and speaking similarly during conversation is a form of synchrony, and is one way in which cognitive information may be transferred nonlinguistically between members of a conversation.

Although many studies have analyzed verbal and nonverbal communication, both separately and together, not enough research explores the relationship between verbal and nonverbal communication in dyadic conversations while incorporating the full dynamics of the communications during the conversations. When studies analyze dynamics, they focus on these communications within a person, or only analyze this relationship within a short time span (i.e., after speaking a few sentences), where between-subject dynamics

might not be noticeable. Furthermore, research that incorporates dynamics has been done mostly outside the field of psychology. The relationships between these dynamics should be further explored, both in a behavioral and quantitative context.

Given the gap in this research, I have therefore performed a set of methods that model the dynamics of facial expressions and nonlinguistic voice within two to four minute dyadic conversations. In this paper, I use active appearance models, short-term Fourier transforms, windowed cross-correlations (WCC), and canonical correlation analysis (CCA) to explore synchrony in these conversations. Furthermore, I explore a novel multivariate time-series method that not only uses CCA to look at the relationship between both a speaker and a listener, but also incorporates the lag between them throughout the conversation. This project thus focuses on the dynamics of the nonlinguistic component of speech through audio clips, and the relationship between facial expressions and the vocal patterns.

### 1.1.1 Overview

I begin by providing background information about communication, conversations, and synchrony, before moving on to how facial expressions and speech work on a physical level and how they are related to each other in communication. For each of these sections, I also operationalize each set of variables by discussing at length the analyses and transformations that were used in order to perform the final canonical correlation analyses. I then describe canonical correlation analyses and windowed cross-correlations and how using them should be beneficial in studying the dynamics of communication and similar types of processes.

## 1.2 Communication and Conversations

Conversation can be defined as the process of communicating between two or more people, and communication to be the process by which cognitive information (thoughts,

ideas, emotions, etc.) is transferred between individuals. When only two participants are in the conversation, the conversation is considered to be dyadic. There are two main components of conversations. The first is the verbal component, which is represented by speech. The verbal aspect of conversation can be further broken down even farther into linguistic and paralinguistic processes (Lieberman, 1975). A linguistic process relates to the semantic content of the speech, while a paralinguistic process relates to other aspects of voice, such as pitch, loudness, and inflection. The other main component of conversations is the nonverbal component, which includes head and body movements, such as head nods, facial expressions, and gestures.

Each component in a conversation leads to communication by providing different avenues for cognitive information can be transferred. For instance, affective information can be transferred through facial expressions and voice inflection, while interpretable thoughts and ideas can be communicated through the linguistic aspects of speech. As each component dynamically changes during a conversation, these changes of the components over time are then interpreted as cognitive information. As a speaker changes their facial expressions, a listener will receive (and presumably understand) these changes as changes in emotion throughout a conversation. Likewise, changes in pitch and loudness of a voice during conversation can also be understood as affective information.

## 1.3 Synchrony and Dynamics

One way in which information is transferred during communication is by using synchrony. Synchrony is the process by which the movements of two bodies in a system match up for some nontrivial period of time, before the movements "break off" and become unique again (Ashenfelter, Boker, Waddell, & Vitanov, 2009). For example, when one person in a dyad smiles, the other person may also return the smile, either with their eyes or mouth. Similarly, when a speaker speaks more quickly for some period of time, the listener,

when he or she then speaks, may also then speak more quickly. These examples can also be seen in the relationship between visual and auditory dynamics, as when a speaker speaks more quickly or loudly, the listener may nod their head more vigorously, as well.

This synchrony can be initiated by one participant in the conversation, and then broken and reinitiated by another participant. The swapping of roles between the leader and follower in this system of synchrony is known as turn-taking (Terven, Raducanu, de Luna, & Salas, 2016; Hung & Gatica-Perez, 2010). Synchrony in conversations has at least two origins. Nonverbal synchrony between two people has been hypothesized to originate in the mirror-neuron system in the brain. Once synchronous movement is realized, various parts of the brain – specifically those that make up the parietofrontal mirror system and the limbic mirror system – activate and continue the symmetry of motor movement (Cattaneo & Rizzolatti, 2009). After one person moves in a way that is independent of the symmetric movement, the symmetry breaks, and the cycle begins again, possibly with the symmetry breaker leading the next symmetric movement. In terms of speech, previous research has shown that the concept of turn-taking is a lot ”messier” than most studies imply – that is, people tend to talk over each other more frequently, as listeners tend to make mistakes in predicting when a speaker will stop talking, and when the listener can start speaking themselves (Heldner & Edlund, 2010; Hung & Gatica-Perez, 2010).

In addition to the overall complexity of the interpersonal dynamics for each dyad, it is also important to note that there is a small delay between the facial expressions and speech patterns of one person, and the facial expressions and speech patterns of the other. Although the movements of people may be synchronized, the movements of one can lag behind the movements of the other. For example, when thinking about how one participant may smile before the other, there is a delay in the onset of the two smiles. On a neurobiological level, this makes sense - it takes some time for the first smile to be seen, to perceive and interpret the action as a smile, and then respond back in kind. In fact, because this delay is expected, when nonverbal delay is manipulated during a conversation, this may

lead to higher anxiety between the participants (Pearson et al., 2008).

Previous research has reported approximately a one second delay between speakers and listeners in terms of both rigid and nonrigid head movement (Ashenfelter et al., 2009; Boker et al., 2009). In terms of audio, there is also a delay between words being spoken, and those words being heard and understood, and a delay can be seen between nonverbal and verbal movement, as well. A study done by Obermeier suggests that the connection between speech and nonverbal gestures is typically understood, at least for native German speakers, when the speech and gestures happen between 120 and 200 milliseconds apart (Obermeier & Gunter, 2015).

## 1.4 Conversations as physical systems

Thinking about this conversation system in the physiological sense, two main aspects of communication are the nonverbal synchronization of movements, and the verbal relationship between a speaker's vocal tract and a listener's ear.

### 1.4.1 Facial Expressions

One of the main functions of facial expressions in conversation is to work in tandem with speech to transfer information between people – in this case, nonverbally. Although many studies have extracted affective information through static facial expressions, it is important to consider the dynamics of facial expressions across entire conversations. (see Krumhuber, Kappas, & Manstead, 2013; Arsalidou, Morris, & Taylor, 2011).

Here a facial expression is defined as a combination of the position and movement of facial features, based on the tightening and loosening of the underlying facial muscles. For instance, a surprised face can be defined as a face where the jaw is slack, and the muscles around the eye are contracted so that the eyes (and sometimes mouth) appear wide open. The dynamics of facial expressions, then, come from the combination of the position

and the movement of these facial muscles over time.

By defining facial expressions in this way, we can then hypothesize how this leads to affective information being transferred between people. As humans develop, we learn how to interpret faces from the dynamics of the muscles, and can then understand the underlying affective information that is represented in a particular face. Because, consciously or otherwise, the relationships between particular facial expressions and their underlying emotions can be acknowledged, we likewise learn as we age how to express particular emotions through our faces, and be able to associate our own facial expressions with the emotions that we feel.

During conversation, then, we can use this knowledge to provide our affect nonverbally to whom we are talking or listening. When participants in a dyad experience emotions, from context given either before or during the conversation, they will likely move their faces to express this affect to the other person. As the faces of participants in a dyad move throughout their conversation, the participants can then use both facial expressions shown to them at any given moment and how these facial expressions move and change over time to interpret how the other person feels at any given moment.

Also, during conversations, the movement of one person's face (typically the speaker) tends to synchronize from time to time with the other person's face. When this happens, the internal association between one's own facial expressions and his or her corresponding affective information means that the affect of both people may synchronize as well. Therefore, as the amount of synchrony between faces increases, empathy is expected to increase, too. Synchrony of nonverbal movement has been detected by using active appearance models (Terven, Raducanu, de Luna, & Salas, 2016; Ashenfelter, Boker, Waddell, & Vitanov, 2009). Gaussian mixture models have also been used to explore rigid head movement synchrony (Xiao, Georgiou, Lee, Baucom, & Narayanan, 2013).

### 1.4.2 Speech

Speech is another important component of synchrony of communication between two people in conversations. Facial expressions provide one way in which cognitive and affective information can be nonverbally communicated, while speech provides cognitive information and affective communication as a part of a verbal stream.

The audio component of human conversations carries two main pieces of information. Linguistic components may contain thoughts and ideas, as well as the speaker's affect. This information is dependent on the definitions and linguistic context of the words being spoken. In addition, paralinguistic information, such as inflections, pitch, and other features of audio, may also contain affective information, which can lead to empathy if properly interpreted.

Biologically, when a listener hears sound from any context, the sound is first transcoded by the ear and then converted into neural signals by the brain. These signals are then interpreted based on the characteristics of the signals, including pitch, or frequency; intenseness, or amplitude; and duration (Darrow, 1990). More specifically, sound is a wave that moves through a medium, typically air, before it reaches our ears (Darrow, 1990). When sound reaches our ears, the movement of the vibrations in the air vibrates our eardrums, which then moves the three bones in the middle ear. This movement then vibrates the fluid in the cochlea, which moves specific hairs inside, depending on the frequencies of the sound waves being heard (de Boer, 1980). The oscillation of the hairs is then converted into electrical signals, which then become pulses in the nerves attached to the hair (Darrow, 1990; de Boer, 1991). Finally, the central nervous system interprets patterns of pulses as sound features and distinguishes between sounds, and the pulses are sent to the auditory cortex for processing (de Boer, 1991; Darrow, 1990).

Speech that a listener hears originates from a speaker's vocal tract. The vocal tract is the physical system containing the lips, tongue, mouth, and vocal cords, as well as the nasal tract (Rabiner & Juang, 1993). These areas are used in conjunction with each other to

produce speech. When a person speaks, the muscles of the face around the mouth contract and relax until they are in a specific position. As the speaker breathes out, air is pushed out of the lungs past the vocal cords and out of the mouth. The speed at which the air travels, the position of the speaker's lips and tongue, and how much their vocal cords vibrate lead to specific phonemes being spoken, which can then be connected together and translated to sound signals to be heard by the listener (Ravindran, Shenbagadevi, & Selvam, 2010; Rabiner & Juang, 1993).

Because speech can be represented as a sound signal, it can be represented as a series of frequencies and amplitudes that change across time. These frequencies depend on variations in how the sound is produced, and individual differences in the structure of the vocal tract. For example, one way in which speech can be sorted is by voiced speech, where the air moving through the vocal tract is affected by the vibration of the vocal cords, and unvoiced speech, where sound is emitted from the vocal tract, but the cords do not vibrate (Yegnanarayana, d'Alessandro, & Darsinos, 1998; Rabiner & Juang, 1993). Due to the movement of the vocal cords, voiced speech tends to be quasi-periodic, while unvoiced speech is aperiodic (Yegnanarayana et al., 1998). Frequencies of speech are also affected by the natural frequencies of each individual's vocal tract. Differences in vocal tract length influence the fundamental frequency of the voice, and therefore frequencies in speech patterns are affected by sex (Honorof & Whalen, 2010). In fact, Honorof and Whalen found that this difference is so great that unless an individual's natural frequencies, or pitch, were close to the "average fundamental frequency" of the opposite sex, listeners could determine the sex of a speaker by their voice alone (2010).

As noted before, synchrony of facial movement can be found as part of dynamics of conversations. Since the movement of the mouth muscles help to define both facial expressions and the sounds that come out of our mouths, then there is a direct relationship between facial muscles near the mouth and specific speech patterns that a speaker makes (Darrow, 1990). In terms of measuring the presence of synchrony in verbal patterns, wavelet trans-

forms have been used to study synchrony in speech (Fujiwara & Daibo, 2016). Synchrony of vocal tract motion has also been found through correlation map analysis, an analysis similar to windowed cross-correlation (Vatikiotis-Bateson, Barbosa, & Best, 2014). From these studies, I argue that synchrony is likely to be present as part of the audio component of conversations as well.

Although this wasn't labeled as such, univariate speech data has previously been analyzed using windowed cross-correlations. Vatikiotis-Bateson used a similar analysis called correlation map analysis that finds instantaneous correlations to model the similarity of motion between two vocal tracts, and found that there was coordination of the different parts of the tract between both participants (Vatikiotis-Bateson, Barbosa, & Best, 2014).

### 1.4.3 Audiovisual relationship

Both audio and visual components of conversations are important to analyze when studying the dynamics of communication. However, while many studies have looked at speech and nonverbal aspects of conversations previously, fewer studies have explored the interaction between them, and fewer still have incorporated the complexity of audio and motion dynamics into these analyses.

As people speak to each other, both verbal and nonverbal components of each person are being analyzed by the other person. These components work in tandem with each other to provide the listener a complete picture of the affective information that is being sent to them. Although much of the information comes from visual input, audio data still plays a part in this transfer of cognitive information. For example, by using echo-state networks, affective information in the form of emotions may be able to be identified from speech signals (Trentin, Scherer, & Schwenker, 2015).

Context is also important for the audiovisual relationship in conversations. Although most of the time, the different components of audio and video provide similar information about the affect of the sender, this is not always the case. If, for example,

someone is visibly upset, but then mentions that he or she is fine, there is a clear disconnect between the linguistic part of the response and the visual cues of the person. Usually in a situation like this, the paralinguistic part of the person's voice will then match up with the visual component, and the person's distress will also be represented by a slower vibrato (Reyland, 2011). Less typically, that person's voice may fail to match the facial expression shown, and the person may look or sound upset, but not both. The mismatch might be thus interpreted as the person showing a fake emotion, or demonstrating masking.

Hani Yehia (Yehia, Rubin, & Vatikiotis-Bateson, 1998) looked into the relationship between the vocal tract and facial expressions on a within-person level. In their study, they use iLEDs on the faces of two persons at separate times. The participants were asked to speak 2-4 sentences in their native language (English and Japanese, respectively) in repetition. The visual and audio data were recorded nonsimultaneously, realigned temporally, and linear estimators were calculated for the visual and audio data. Although the study focuses on the within-person relationship between verbal and facial behavior, and the time span in which the participants were speaking is much shorter than can be found in most everyday conversations, this study still provides a possible window into the audiovisual relationship found in conversations between people.

Other studies have also reported associations between speech and facial expressions. For example, in one study, an emotion recognition system and support vector machine classifiers were used to explore how both speech and facial expressions work together to output affective information on an individual level (Busso et al., 2004). In a study by Hung, it was reported that there is a direct relationship between audiovisual nonverbal behavior and small group cohesion (Hung & Gatica-Perez, 2010). Kim, Cvejic, and Davis found that eyebrow movement, and rigid head nodding after some delay, both occur while speaking (Kim, Cvejic, & Davis, 2014). The connection between vocal patterns and facial expressions can be seen through lip smacking in nonhuman primates, as well (Ghazanfar & Takahashi, 2014).

In terms of audiovisual synchrony, by modeling speech and facial expressions as "locally Gaussian distributions," and measuring the amount of mutual information between them, synchrony can be found within a person between his or her face and speech patterns (Iyengar, Nock, & Neti, 2003). Outside of speech, listening to certain music or television shows can also lead to rhythmic synchronous body movement to the sound from the music or TV (Sejdić, Jeffery, Kroonenberg, & Chau, 2012).

## 1.5 Rationale

Although audiovisual data has been modeled and analyzed as univariate structures before, the simplicity of this type of data used may not have inherently encompassed the full complexity of the underlying system found in dyadic conversations. Various studies have reported that rigid head movement are able to be reduced to four main dimensions – namely, vertical head nods, horizontal head shakes, oblique head tilts, and forward and backward head movement. Most nonrigid head movement can also be expressed as an eight-dimensional structure, due to the capabilities and limitations of the possible movements of facial muscles (Boker et al., 2011). Due to the vast amount of components that make up the vocal tract and contribute to the formation of speech, including the movement of tongue and lips and the vibration of the vocal cords, the resulting vocal patterns require at least four or more dimensions to fully model properly (Vatikiotis-Bateson, Barbosa, & Best, 2014). Therefore, by viewing audiovisual data as multidimensional structures, more of the dynamics and complexity needed to understand and explore communication processes such as synchrony across time may be more comprehensively captured.

I therefore propose using a novel, windowed variant of canonical correlation analyses that would be able to handle the multidimensional structure of the facial and speech data while incorporating the dynamics found in dyadic conversations. I also propose using windowed cross-correlations as a univariate analogue to the method of windowed canon-

ical correlation analyses. To show how using these methods work, I begin by explaining where the original video and audio clips originate, and how to decompose the clips into their multidimensional components using active appearance models and short-time Fourier transforms, respectively. Then, I will explain how I explored the relationships between speakers and listeners in dyadic conversations by performing windowed cross-correlations and windowed canonical correlation analyses.

# 2. Methods

Previous research has estimated synchrony of facial expressions and the presence of turn-taking in dyadic conversations from windowed cross- correlations and factor analyses that incorporated constant and variable time delays. I have incorporated audio and video data from preexisting dyadic conversations into two sets of analyses to estimate synchrony and turn-taking in conversations when both audio and visual data are in the same models.

## 2.1 Participants

In the experiment where the video clips originated, 106 undergraduate students from a southeastern university who were previously unacquainted entered booths in separate rooms, and were asked to have unscripted, two-to-four minute conversations with each other over closed-circuit video (62 Females, 39 Males, 5 unreported). The experiment was broken into two studies, and every participant only appeared in one of the two studies. The participants were volunteers from the psychology department participant pool, and were given course credit for completing the experiment. As part of pre-selection criteria, all participants were fluent in English and were not aware that their motion was being tracked. The participants only saw each other as video projected on screens in their booths. Neither participant knew each other before the experiment began; thus, the participants were given two minutes to introduce themselves across the video feed and get to know each other before they were given the prompt for the main conversations.

In each booth, the video feed of the opposite participant was projected onto a screen, and the participants were told that they will see video of either the participant from the shoulders up, or just the face of the other participant cut out on the screen. The researchers gave the participants hats with three plastic screws to wear to provide visual anchor points

for the face tracking software to more easily track the face. The participants were told that the researchers were "measuring magnetic fields during conversation" to prevent the participants from knowing the true purpose of the hat during the experiment (Boker & Cohn, 2009). All participants read and signed Institutional Review Board-approved informed consent forms, and were debriefed at the end of the experiment (Boker et al., 2011).

## 2.2 Experiment

For each emotion-memory conversation, one of the participants was given the prompt to "describe a memory where you felt" one of four emotions – happiness, sadness, surprise, or disappointment. The person who was describing the memory was designated the "speaker," while the other person was designated the "listener." However, neither participant was forbidden from speaking to each other during the conversations. For each emotion, one participant was first designated as the speaker, and then the other. For the first study, each pair of participants had eight conversations, and were given one of the four emotions for the prompt. In the second study, each pair of participants had six conversations, and was given either "happy," "sad," or "disappointed" for the prompt. All conversations were two to four minutes long ($M_{seconds}$ = 127.55).

Video and audio was recorded in all conversations during the experiment. Cameras in the booths recorded video at a frame rate of 29.97 frames per second. Earthworks directional microphones recorded the participants' speech, and each participant wore headphones to be able to hear the other person (Boker et al., 2011). The resulting audio clips were in 16-bit PCM WAVE format, with a sampling rate of 44100 samples per second. The video feeds were synchronized to the nearest .1 millisecond by using a MOTU V4HD digital recorder and a GPS master clock (Boker et al., 2011). Due to the video and audio signals processing and traveling at different speeds between the two booths, there was a 33 millisecond delay between the audio and video feeds (Boker & Cohn, 2009). To cor-

rect for the lag in video, a Yamaha 01V96 mixer delayed the audio for each person by 33 ms to match the video (Boker et al., 2009, 2011). The recorded video clips depicted each participant separately; that is, for every conversation, two video clips were made. For the first study, each participant's audio was recorded as separate audio files, while in the second study, both participants' audio were saved as the left and right channels of a single conversation-level audio file. Because all audio and video streams were synchronized when recording the conversations, any lag noted between the speaker and listener in the video and audio clips is naturally performed by the participants themselves.

## 2.3 Data

### 2.3.1 Facial expression data

#### Active Appearance Model

To quantify the dynamics of faces to explore the synchrony of facial expressions between dyads, active appearance models have been used to track the deformations of the face (or facial expressions) across 28 dyads, or pairs of participants, from the conversations from both studies in the original experiment. An active appearance model is a model that estimates the shape and appearance of a particular object as it changes over time (Cootes, Edwards, & Taylor, 2001). The algorithm used in this process uses principal component analysis (PCA) to help match a series of points on a deformable mesh with the major features of the face on each frame of a video clip. To do this, the model is manually trained on 20-30 manually tracked frames to match appropriately the various parts of the face. These frames were chosen to maximize the variance of facial expressions and rigid head positions, which would minimize any errors from mistracking the face. The active appearance model algorithm uses the trained frames and pixel differences from the other image frames and deviations from the built model to estimate where the points on the mesh

should be on the other frames (Cootes, Edwards, & Taylor, 2001). Figures 2.1 and 2.2 show the program used to construct the models, and the points from a similar frame plotted in R. The point coordinates of the model from the tracked frames were used as data to represent facial expressions in the proposed analyses.

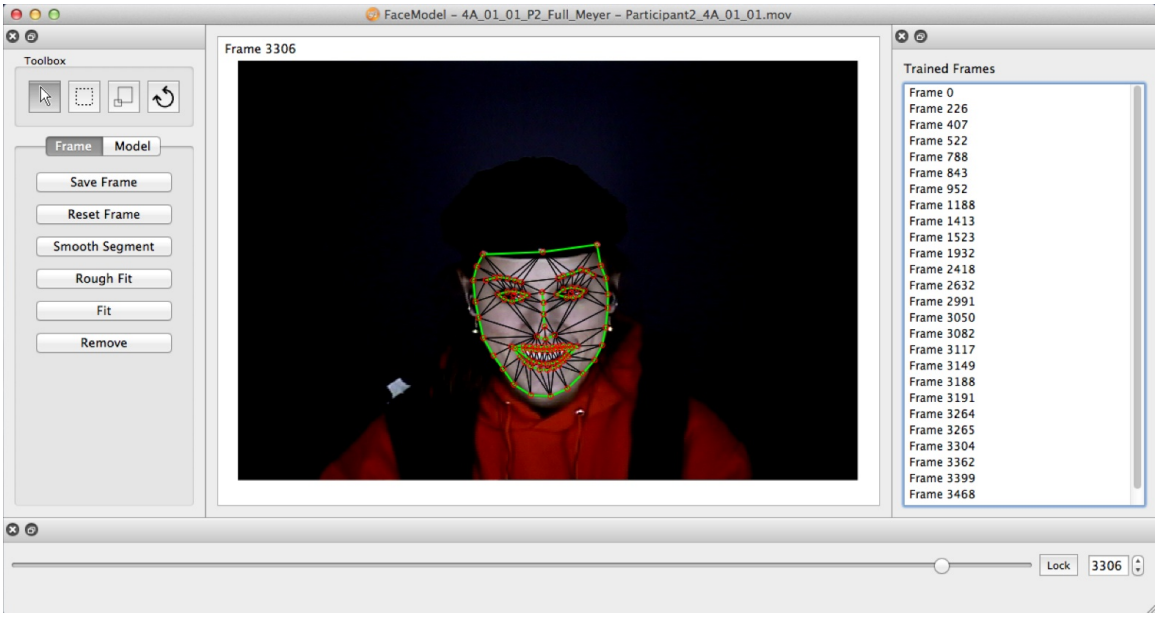



*Figure* 2.1: The GUI frontend of the program used to construct the active appearance models and track the faces from the recorded conversations. The participant pictured is the designated listener from the first conversation pair, and the current frame is frame 3306 from the first conversation. In this conversation, the speaker was asked to describe a memory when they felt happiness.

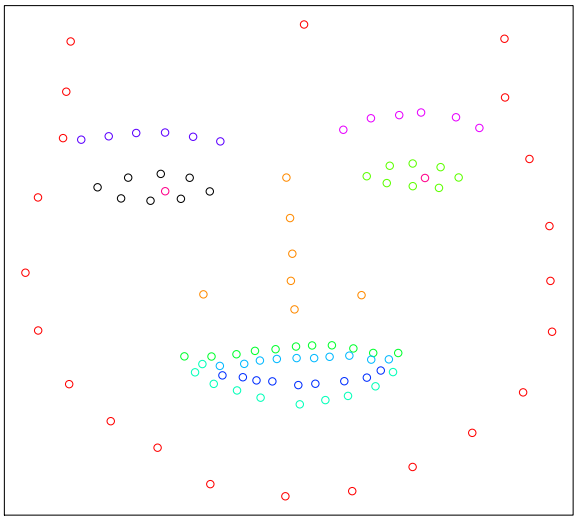


*Figure* 2.2: A plot of the mesh from the active appearance models, in the form of colored points in the shape of the face. The face depicted is the first video frame of the speaker from the first conversation. Each major facial feature is represented as a different color.

To analyze the dynamics of these point coordinates, data matrices have been constructed from the outputted point files. The data matrix for each person is

$$C_{i,R,points} = \begin{bmatrix} X_{1,i,R,1} & Y_{1,i,R,1} & X_{2,i,R,1} & Y_{2,i,R,1} & \cdots & X_{P,i,R,1} & Y_{P,i,R,1} \\ X_{1,i,R,2} & Y_{1,i,R,2} & X_{2,i,R,2} & Y_{2,i,R,2} & \cdots & X_{P,i,R,2} & Y_{P,i,R,2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{1,i,R,F} & Y_{1,i,R,F} & X_{2,i,R,F} & Y_{2,i,R,F} & \cdots & X_{P,i,R,F} & Y_{P,i,R,F} \end{bmatrix} \tag{2.1}$$

where each row represents $P$ horizontal (x) and vertical (y) coordinates of the vertices of the facial mesh for one video frame of one person from the conversation. Here, $i$ represents which conversation, or pair, the matrix originates; $R$ represents the designated role of the participant: either a speaker (S) or a listener (L); and $F$ represents the number of video frames in the original video clip.

Active appearance models have been used in several studies when studying facial expression dynamics. Girard and his collaborators used a system that required manual coding by researchers – the facial action coding system (FACS) – and active appearance models (AAM) as an automatic tool to study how facial expressions and rigid head movement relate to depression symptoms. The study reported that not only was this relationship significant, but that AAM did as well in showing this relationship as FACS. (Girard et al., 2014). Although active appearance models do not capture every minute detail of the faces being tracked, a study by Theobald reported that facial expressions are understood between people even when participants only see commonalities of the physical face features (Theobald et al., 2009). In a study by Boker that used AAM to construct face avatars similarly to the proposed point coordinate data, participants naive to the design did not notice that the avatars were distinguishable from a video stream of the confederate (Boker & Cohn, 2009). From this evidence, the point coordinate information is likely to be sufficient for estimating synchrony.

**Extracting nonrigid head components**

When modeling the dynamics of facial expressions by using point position data from active appearance models, the point position matrices must first be decomposed, and the nonrigid head components need to be extracted. From previous research, when performing singular value decomposition (SVD) on the point position data and looking at the dimensionality of this data, the first four head movement components tend to represent rigid head movements. Once these are removed, the next eight to twelve head components are associated with nonrigid head components, such as facial expressions (Boker et al., 2009). Therefore, we can focus on the positions and movements of facial expressions by focusing on these components of the matrix.

The equation for performing SVD on the point coordinate matrix is

$$\mathbf{U}\Gamma\mathbf{V}, \tag{2.2}$$

where **U** and **V** are the left and right singular vectors of the point coordinate matrix, respectively, and $\Gamma$ is the diagonal matrix of singular roots. Once the rigid head components are removed, the matrix can be reconstructed using the same method.

### 2.3.2 Speech data

To model the dynamics in the audio, preprocessing was performed on the audio clips in order to ensure that only the speech of the speaker and listener are being analyzed. First, the audio signal was centered at the mean, otherwise known as removing the DC offset. This was to reduce the effect any noise outside of the participants' voices, and any other audio artifacts, may have on the analysis. This was done by subtracting the mean amplitude for each audio channel in the audio clips from the respective audio channel.

Furthermore, assuming a decibel level of 50 dB, which is the average intensity level for normal conversations (Darrow, 1990), humans can typically hear frequencies between around 25 Hz to almost 20 kHz (de Boer, 1980). Also, it is noted that the noise component of audio recordings of speech are contained in higher frequencies, typically above 3-4 kHz (Yegnanarayana, d'Alessandro, & Darsinos, 1998). Amplitudes that are outside the range of human speech (between 50 Hz and 8.192 kHz) were therefore removed by removing the frequencies directly after the signal is decomposed.

#### Short-time Fourier Transform

In order to look at synchrony of pitch (or frequency) and loudness (or amplitude) between the participants in the conversations, a short-term Fourier transform, or STFT, was used to convert our audio signals to the frequency domain. STFT is a variant of a

Fourier transform, which is a method that converts a raw audio signal into the frequency domain from the time domain (Aamir & Maud, 2007). From the results of the Fourier transform when performed on audio, a spectrogram can be produced, where amplitudes are output based on frequency. While a Fourier transform performs this conversion across an entire audio sample, STFT does this over sections, or windows, of the audio. Using STFT therefore provides a way to look at changes in frequencies and amplitudes across time and thus changes in speech dynamics.

This process was performed by taking windows, or sections of the original audio wave signal, and performing a linear transformation of the signal within each section, such that each resultant signal has a basis of sinusoidal functions, which represent frequencies. Since Fourier transforms output a reduction of signal information near the edges of the data that's being transformed, this can be rectified by overlapping each window with the previous one. This overlap, then, adds redundancy to the information that would otherwise be lost near the edges of each window (Mowlaee, Saeidi, & Stylianou, 2016).

Although STFT has been ultimately chosen as the transformation of choice for analyzing the audio clips, other transformations were considered. One popular alternative to STFT is the wavelet transform. The wavelet transform has been used in the past to study synchrony of speech (Fujiwara & Daibo, 2016). However, it has been found that orthogonality of speech signals increases when using wavelet transform over STFT (Tinati & Mozaffary, 2005), which is useful for speech separation (i.e., between multiple voices), but less so when performing canonical correlation analysis. In a study by Fan, a comparison between wavelet and STFT was made for "single channel speech signal noise reduction," which showed that STFT was preferred over wavelet transforms in terms of reducing noise (Fan, Balan, & Rosca, 2004).

STFT was used in the current proposed study to convert raw audio data from the video clips into the frequency domain for fixed intervals of time across the conversations. Since STFT operates over both the time and frequency domain, it is important to find the

optimal window size that maximizes the information of both of these domains (Paliwal & Alsteris, 2003). Similarly, to prevent information being lost on the edges of the transformation, the best overlap of windows to use must be chosen. Since the main comparison being discussed is between speech dynamics of conversations and facial expressions, a window size for the STFT that can easily match up with the smallest time frame of the facial expression data must also be considered: 1 video frame, or $\frac{1}{29.95} = 33$ ms.

One possible window size typically proposed is a window duration of less than 100 ms. This is because using this window size satisfies the assumption that Fourier analyses should only be performed on stationary or quasistationary data. Since speech is at least quasistationary at this small time frame, using STFT with this window size, there is less of a concern that this assumption was violated (Rabiner & Juang, 1993). In addition, Paliwal used a consonant recognition task and an automatic speech recognition task to show that when audio is recreated from the short-time magnitude spectrum, speech was most understood when a window duration of 15-35 ms was used (Paliwal, Lyons, & Wojcicki, 2010).

As it is standard for speech data, a Hamming window with an overlap of 50% is typically used when performing STFT (see Avargel & Cohen, 2007; Paliwal et al., 2010; Stanković, Orović, Stanković, & Amin, 2013). The equation for this window is

$$W(n) = 0.54 - 0.46 \cos \frac{2\pi n}{N-1}, \tag{2.3}$$

where n is between 0 and one less than the number of samples per window (Muda, Begam, & Elamvazuthi, 2010).

From these recommendations and constraints, an STFT window size of 16 2/3 ms, or about 736 audio samples, was chosen. This window size was chosen because not only would it be within the recommended range of window sizes for this type of data, but it would also equal roughly 1/2 a video frame, which, with the chosen overlap, would mean that there would be 4 audio observations per video frame of data. Having the number of

audio observations be divisible by the number of video observations allowed the speech and visual data to be compared more easily.

The data matrix, then, is

$$C_{i,R,freq} = \begin{bmatrix} A_{1,i,R,1} & A_{2,i,R,1} & \cdots & A_{368,i,R,1} \\ A_{1,i,R,2} & A_{2,i,R,2} & \cdots & A_{368,i,R,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,i,R,4*F} & A_{2,i,R,4*F} & \cdots & A_{368,i,R,4*F} \end{bmatrix}, \tag{2.4}$$

where each row represents the amplitudes corresponding to 368 frequencies from the STFT for 1/4 of a video frame (1/120 second) of one person from the conversation. After removing the frequencies outside of the boundaries of human hearing, each matrix had 136 columns, or variables, that corresponded to the audible frequencies. Here i represents the conversation, or pair, from which the matrix originates; R represents the designated role of the participant: either a speaker (S) or a listener (L); and F represents the number of video frames in the original video clip.

## 2.4 Analyses

After the facial expression and audio data were constructed and preprocessed, I performed two sets of analyses. The first set of analyses modeled the synchrony of facial expressions between two participants in a dyadic conversation, and the synchrony of speech between the same conversation dyads. This included comparing the facial expressions of one participant to the facial expressions of the other participant, and comparing the speech of one participant with the speech of the other participant. The second set of analyses was designed to explore the relationship between the speech from one participant in the dyadic conversations with the facial expressions of the other participant. This included pairing the designated speaker's speech with the designated listener's facial expressions

and the designated listener's speech with the designated speaker's facial expressions. Based on these combinations of pairs, this ultimately led to four separate analyses. Table 2.1 summarizes what analyses were performed.

| Data | Speaker Faces | Speaker Speech |
|---|---|---|
| Listener Faces | Between Faces | Speaker Speech/Listener Face |
| Listener Speech | Speaker Face/Listener Speech | Between Speech |

Table 2.1: Summary of analyses being performed.

### 2.4.1 Windowed Cross-Correlation

For each set of analyses, windowed cross-correlations (WCC) were performed on the data. Windowed cross-correlations are cross-correlations between pairs of sections (or windows) of two time series. The pair of windows can either be next to each other, or farther apart. The difference in the pair of windows is called "lag," and represents the lead-lag time, or temporal distance, between the two time series. For each pair of windows, the cross-correlation between them is first calculated by finding the correlation between the two windows. Then, for some lag $\tau$, one of the two windows is shifted down $\tau$ rows of the corresponding time series, and the cross-correlation is calculated for the new pair of windows. This process then continues until the distance between the windows is equal to some maximum lag $\tau_{max}$, and then the first window is shifted back to its original position while the process repeats for the opposite window. Once both windows have been shifted to the maximum lag, the position of the lagged window is reset to the same position as the nonlagged window, both windows are shifted by some window increment $\omega$, and the process begins anew until the windows have been incremented across the entire time series (Boker, Xu, Rotondo, & King, 2002). Figure 2.3 provides a visual summary of this analysis.

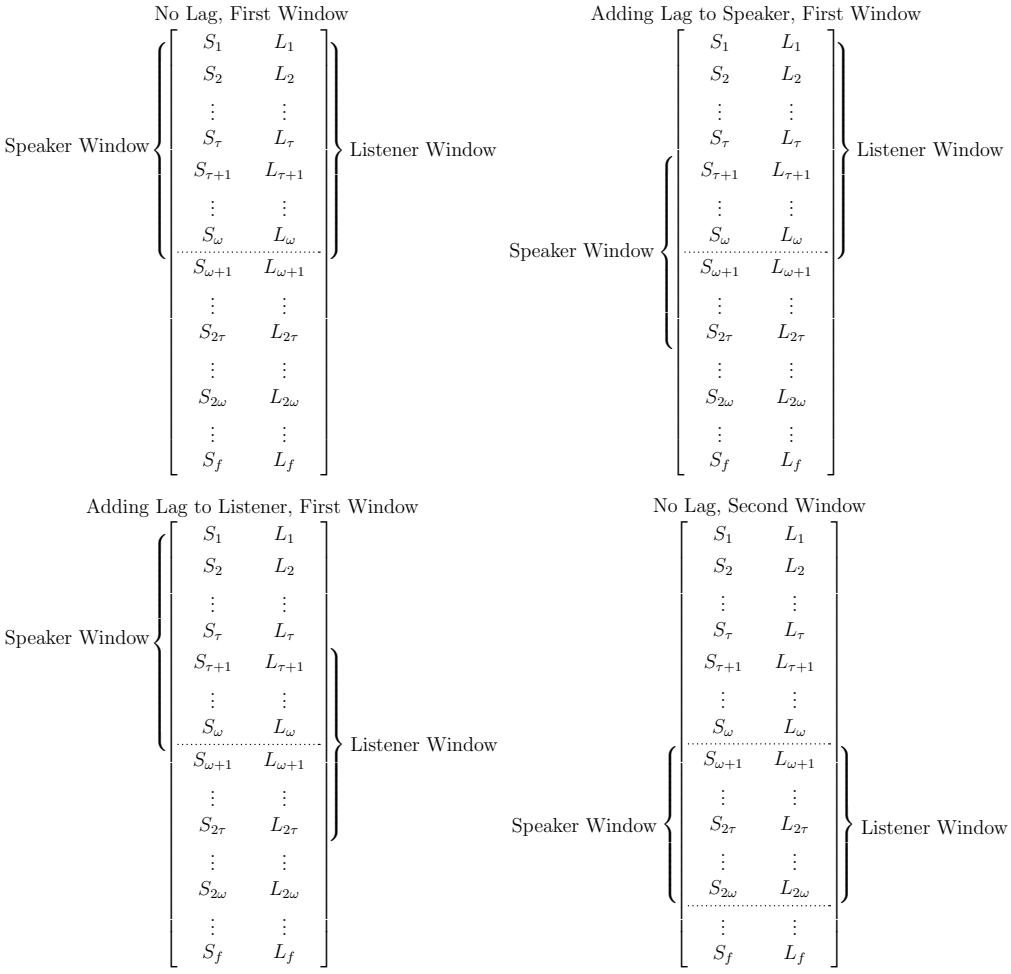



*Figure* 2.3: Visualization of windowed cross-correlation analysis. In each matrix, each row represents either one video frame or one STFT window, and each column represents one of the two time series being analyzed – either the root mean square (RMS) point coordinates or the RMS amplitude for a speaker ($S$) and listener ($L$) of a conversation dyad. As part of the process of calculating the cross-correlations, windows as subsets of the two time series are first extracted before lag is taken into account (top-left matrix). Once cross-correlations are found for those windows, either the speaker window (top-right) or listener window (bottom-left) is shifted a certain number of rows, and the cross-correlation is found for the new windows. Finally, once the cross-correlations are found for all lags of the given windows, both windows are incremented (bottom-right), and the process continues until both windows reach the end of the two time series.

After every cross-correlation was calculated for each window and each lag, a peak-picking algorithm was used to determine the peak, or maximum cross-correlation for each set of windows. For each pair of windows, the peak cross-correlation was found by first

smoothing the cross-correlations with a quadratic loess function. Then, beginning with the lags closest to zero lag and increasing to the most extreme lags, a local search region was used to find the highest value cross-correlations that are "monotonically decreasing on either side of" said correlation (Boker et al., 2002). The maximum cross-correlation that met this criterion was then chosen as the peak value, and the corresponding lag was the peak lag for the set of windows. To remain consistent with previous research on the facial expressions of the same video clips, a 10-15 second window, a maximum lag of five seconds, window and lag increments of at most 1/3 second, and a loess function span ($pspan$) of .1 and a local search region size ($L_{size}$) of 4 units was used.

After this process was complete, all cross-correlations from the windowed cross-correlation analyses were plotted as heatmaps, with the y-axis being each lag, and the x-axis being the time across the time series. On each heatmap, a color of yellow represents a strong positive correlation, and a red color represents a strong negative correlation. The peak lags that correspond to a maximum positive correlation for each window were also plotted as a black line across the graph.

If the peak cross-correlations, represented by the black line in the plots, have several periods of time in which the line is relatively stable, before a large change occurs in the lag, then these stable periods would be seen as times in which the participants are demonstrating symmetry, and the jumps can be considered points in which this symmetry breaking. If this symmetry formation and breaking occurred, we can say that synchrony occurred during the conversation over short intervals. Also, if there is a change in sign of the peak lag, then depending on the sign change, there are times in which the speaker is leading in the synchronous facial feature movement (i.e., the positive lag), and other times in which the listener is leading (the negative lag). This change in sign can therefore be attributed to turn-taking in the conversation.

Windowed cross-correlations were used in the proposed analyses to show one way in which synchrony between dyads can be estimated. To do this, the root mean square

(RMS) point positions and RMS amplitude were calculated for every row of the facial expression and STFT data matrices. Each RMS point position represented a frame of video (1/30th of a second), and RMS amplitude represented one audio sample (1/120th of a second). Windowed cross-correlations were then performed on either the RMS point positions or the RMS amplitudes for each dyad pair. All windowed cross-correlation analyses output a list of peak lags and WCC plots that showed the overall peak lag and correlations across the windows and demonstrated whether turn-taking was happening between participants.

Because each frame of video from the conversations corresponds to one row of point positions and four rows of audio amplitudes, the point positions were upscaled for the WCC that compared the RMS point positions of one participant and the RMS audio amplitudes of the other. To do this, a linear spline interpolation was performed on the point position data before the root mean square point positions are calculated, and intermediate rows of point positions were constructed for each 1/4 video frame (1/120th of a second). Each row in the upscaled point position matrices thus matched up with the corresponding row of a STFT data matrix.

First, WCC was performed on the RMS audio amplitude from the amplitude spectrum between participants in each dyad to find the peak lag between audio in participants throughout the conversation (Between Speech). WCC was then performed on the RMS point positions of each listener and the RMS audio amplitude from the STFT matrices of each speaker in the dyad (Speaker Speech/Listener Face). WCC was also performed on the RMS audio amplitude of the speaker, and the RMS point positions of the listener, to test the peak lag between audio and video throughout the conversation (Speaker Face/Listener Speech). Finally, WCC on RMS point positions between the speakers and listeners (Between Faces) was performed as well. A list of peak lags and WCC plots of turn-taking between participants was then output from the WCC.

Windowed cross-correlations are useful when exploring dynamics because they can show not only peak correlations for each section of pairs of time series, but they can also

show how far behind one time series is to another when the cross-correlation between the two sections of time series is highest. For example, when looking at time series from two people, turn-taking can be seen by observing increases, decreases, and flips in sign of the peak lag between the two people. This has been seen previously with both rigid and nonrigid head movements, among other motion dynamics in human dyads.

### 2.4.2 Windowed Canonical Correlation Analysis

After performing the WCC analyses, I performed canonical correlation analyses on the data. Canonical correlation analysis (or CCA) is a multivariate approach that is used to find the maximal relationship between two sets of variables, say $X_1$ and $X_2$. This method is done by finding linear combinations of $X_1$ and linear combinations of $X_2$ such that the correlations of pairs between the sets of combinations are maximized (Iaci, Sriram, & Yin, 2010). These pairs of linear combinations are defined as canonical variates, or pairs, and the higher the total correlation between significant canonical pairs, the more similar the variables in each set are to each other.

Once these canonical correlations were found, each canonical dimension was tested by calculating Wilk's lambda to test if they are significant. Let $M_1$ and be a matrix of size n x $c_1$, and $M_2$ be a matrix of size n x $c_2$. Then, if $r_{a,b}$ represents the canonical correlations of variables a in $M_1$ and b in $M_2$, the cumulative products of $1 - r^2_{a,b}$ are calculated. These products represent Wilk's lambda for the canonical pair representing variables a and b. Once these are calculated, the equation for the F statistic corresponding to the Wilk's lambda $W_{a,b}$ is

$$\frac{(1 - W^x_{a,b})d_l}{(W^x_{a,b})d_u} \tag{2.5}$$

where $x$ is

$$\sqrt{\frac{c_1^2 + c_2^2 - 5}{c_1^2 * c_2^2 - 4}}. \tag{2.6}$$

This F statistic would then have an upper degrees of freedom of $d_u = c_1 c_2$ and a lower degrees of freedom of

$$d_l = \frac{2n - 3 - (c_1 + c_2)}{2x} - \frac{c_1 c_2}{2} + 1. \tag{2.7}$$

This F statistic can then be tested for significance (Afifi, May, & Clark, 2003).

CCA is useful when exploring dynamics because the synchrony of the dyads can be expressed as two sets of variables - one from the speaker and one from the listener. As mentioned before, facial expressions can be expressed as x and y coordinates of a 99 point two-dimensional mesh across a face, constructed from an active appearance model, and speech from the conversation can be expressed as a multivariate frequency matrix that is constructed from performing STFT on audio clips recorded from the conversation. Using these point coordinate and audio frequency matrices, I then used CCA to examine the co-variance between a speaker's and listener's facial point coordinates and audio frequencies. The greater the total canonical correlation obtained, the more similar the coordinates and frequencies are between the speaker and listener, which then means that the speaker and listener are acting in similar ways, both visually and aurally. Likewise, the proportion of variance explained by the significant canonical pairs from the CCA results can also provide an indicator of how similar the speaker's and listener's faces move, and how similarly they speak. If there are a high number of significant canonical pairs, then the dimensionality of the synchrony between people is high, and is represented by multiple significant canonical pairs. Likewise, if the proportion of shared variance of each canonical pair is high, then the synchrony should be high as well, and may be represented by fewer canonical pairs.

Because time delays exist between people during conversations, it is important to

include a time delay in our model. These time delays can be seen in the windowed cross-correlations in the form of peak lags. For this analysis, a variable time delay was implemented. As mentioned previously, one aspect of speaker-listener dynamics in dyadic conversations is the process of turn-taking. Thus, the role of who is leading and who is lagging can be expected to switch within conversation. Changing lags can be emulated as part of a novel take on multivariate analyses: by adding artificial lag to sections of data rather than the entire conversation, the movements of the two participants should be matched up even more.

To implement the variable time delay, submatrices with the same number of columns as the original matrices, and rows corresponding to a temporal window, were extracted similarly to extracting the windows for the WCC plots. The matrices chosen for the analyses were the STFT matrices from the designated speaker and listener from each conversation dyad to represent the paralinguistic speech of the designated speaker and listener, and the point coordination matrices to represent the facial movement of the designated speaker and listener. The window sizes were chosen with both consideration to smoothing and proper model specification in mind; I therefore chose a window size of 15 seconds, or 450 video frames, so that there would be at least 450 observations for each analysis, and more observations than variables (198 for the point coordinate matrices for each participant, 136 for the STFT matrices). Similarly to the WCC, submatrices were first extracted from the top of the original matrices. Then, to simulate lag, one of the windows was shifted down by the number of rows equivalent to 1/3 second, and the new submatrices were extracted. Once all pairs of submatrices have been extracted for all possible lags up to the maximum lag of 5 seconds, both windows were shifted down, and the algorithm continued until the windows reached the end of the conversation. For each pair of submatrices, canonical correlation analyses were performed, and the number of significant canonical pairs and the variance explained by them were then found for each analysis. Figure 2.4 summarizes this process. Descriptive statistics tables, a figure of the distribution of significant canonical pairs, and

a figure of the distribution of proportion of variance explained were constructed from the results of these analyses.

For the first set of canonical correlation analyses, the point coordinate matrix of each designated speaker was matched with the point coordinate matrix of the corresponding listener from the conversation (Between Faces), and the STFT matrix of the designated speaker was matched with the STFT matrix of the corresponding listener (Between Speech). The STFT matrix of the speaker was then paired with the point coordinate matrix of the listener (Speaker Speech/Listener Face), and the point coordinate matrix of the speaker was paired with the STFT matrix of the listener (Speaker Face/Listener Speech) to test how the speech of one participant synchronizes with the facial expressions and movement of the other participant. Similarly to the WCC, the point position data was upscaled prior to the Speaker Speech/Listener Face and Speaker Face/Listener Speech analyses by performing a linear spline interpolation to construct intermediate rows of point positions for each 1/4 video frame (1/120th of a second). Again, each row in the upscaled point position matrices thus matched up with the corresponding row of a STFT data matrix.

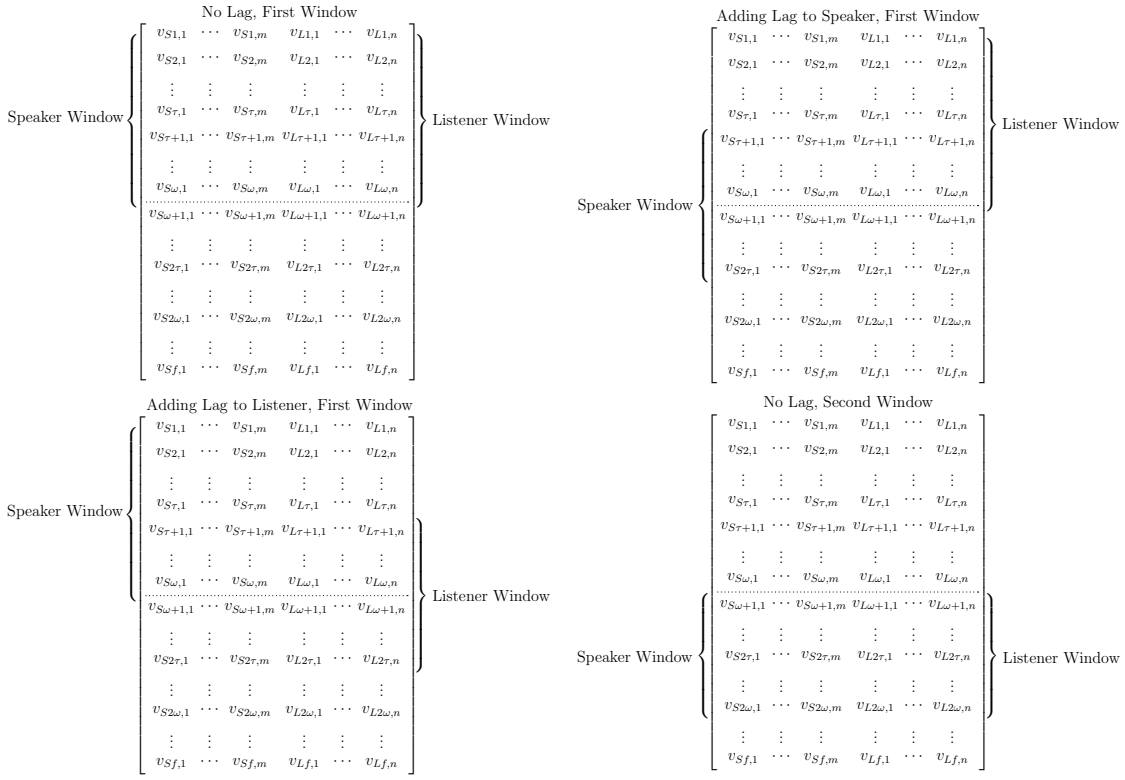



*Figure* 2.4: Implementation of variable time delay into a canonical correlation analysis. Each row in the matrices represents either one video frame or one STFT window, and each column represents a variable in the combined conversation matrices – either an x or y point coordinate ($m$, $n$ = 198) or a frequency from the STFT ($m$, $n$ = 136) for a speaker ($S$) and listener ($L$) of a conversation dyad. Similarly to the method of finding windowed cross-correlations, submatrices are first extracted from the speaker and listener matrices before lag is taken into account (top-left matrix). Once the CCA is ran on these submatrices, either the speaker window (top-right) or listener window (bottom-left) is shifted a certain number of frames, and a CCA is run on the new windows. Finally, once CCA is run for all lags, both windows are incremented (bottom-right), and the process continues until both windows reach the bottom of the combined conversation matrix.

If the CCA result in high correlations (>.900) between the facial expression data or the speech data of the pairs, matrices were created for each participant at this point that only consist of the first 12 left singular vectors of each point coordinate and STFT matrix. This number of left singular vectors was chosen based on previous research. Typically nonrigid head movement can be represented by the first seven to eight components (Boker et al., 2011) and speech can be represented by at least four components (Vatikiotis-Bateson,

Barbosa, & Best, 2014). The CCA was then performed on the left singular matrices, and results from both sets of analyses were reported.

Although canonical correlation analysis has not been used yet in this way to explore synchrony of speech and facial expressions together between people in dyadic conversations, CCA has been used previously in studies looking at speech patterns and facial expressions separately. For example, a study by Melzer used an extension of CCA called kernel canonical correlation analysis (KCCA) as an alternative to PCA when working with appearance models (Melzer, Reiter, & Bischof, 2003). KCCA was also used in a study by Zheng to perform identification of facial expressions (Zheng, Zhou, Zou, & Zhao, 2006). CCA has been used in association with a technique called multimodal fusion to help identify and separate the audio of individual speakers in conversations, as well (Sargin, Yemez, Erzin, & Tekalp, 2007).

### 2.4.3 Surrogate Analysis

Once the time delays are included in the model, the number of significant canonical pairs and corresponding variance explained were used as a possible measurement of how much synchrony there is between the speaker and listener in the conversation. However, as the distribution of significant canonical pairs is unknown for this type of data, traditional parametric tests can't simply be performed to test whether the number of significant canonical pairs found are significantly different than chance. To circumvent this issue, surrogate analyses were used in this project as a substitute to answer the same question.

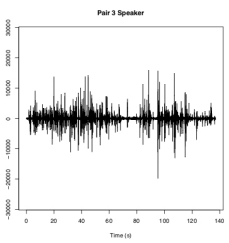
To test the analyses between the speakers and listeners, the surrogate analyses took every designated speaker from the dyads, and paired them with every other designated listener besides the listener from the original pair. The analyses performed on the original dyads were then performed on the new combinations of face and speech data from the speakers and listeners, and the results from the analyses for the original dyads and the surrogate pairs were compared. This comparison should test whether any possible synchrony

found between the movement and speech of the participants of the original pairs is due to the actual conversations, or due to the commonalities found in movements and speech of people independently of interacting with someone in a conversation. Distributions of the number of significant canonical pairs and the corresponding variance explained were plotted for both the original dyads and the surrogate pairs to compare the results of the CCA between the surrogates and the original dyads.

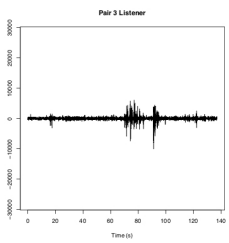
# 3. Results

## 3.1 Waveforms and Spectrograms

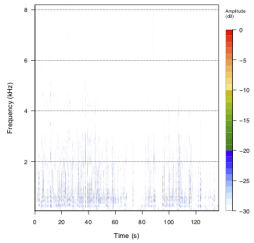
In addition to descriptive statistics and distribution plots that represent the overall results of the WCC analyses and CCA across all the pairs, four pairs were chosen that were representative of the complete 28 pair dataset to best illustrate the results of the analyses on the pair level. These pairs were chosen based on the proportion of time the designated speaker and designated listener each spent talking. Waveforms and spectrograms of these pairs are shown in Figure 3.1. In pairs 3, 6, and 14, the designated speaker was speaking throughout, while neither the speaker nor the listener spoke for a nontrivial amount of time in pair 22. The amount of time the designated listener spoke was different in all four representative pairs, ranging from not at all (pair 14) to sometimes (pairs 3 and 22) to throughout the conversation (pair 6).



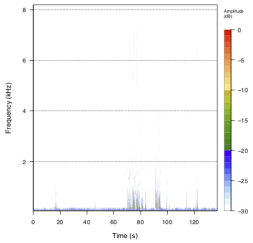

(a) Waveform of Speaker, Pair 3



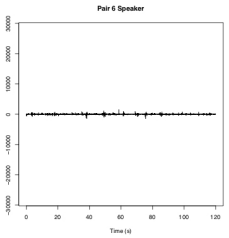

(b) Waveform of Listener, Pair 3



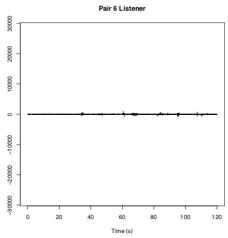

(c) Spectrogram of Speaker, Pair 3



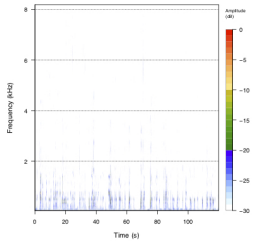

(d) Spectrogram of Listener, Pair 3



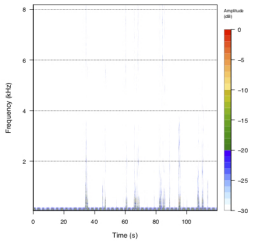

(e) Waveform of Speaker, Pair 6



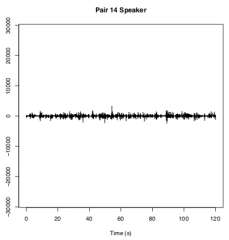

(f) Waveform of Listener, Pair 6



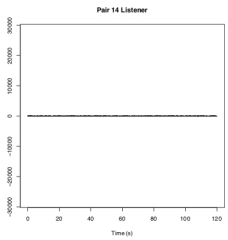

(g) Spectrogram of Speaker, Pair 6



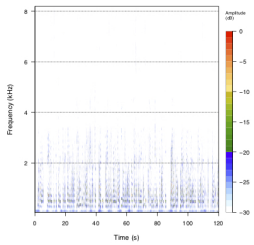

(h) Spectrogram of Listener, Pair 6



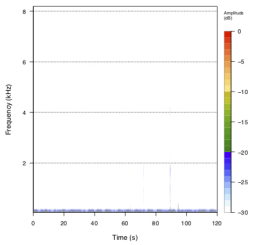

(i) Waveform of Speaker, Pair 14



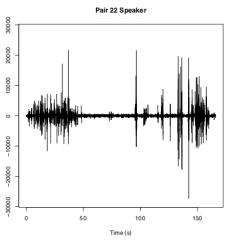

(j) Waveform of Listener, Pair 14



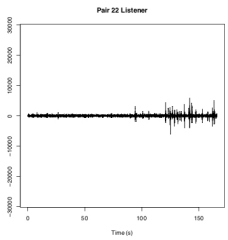

(k) Spectrogram of Speaker, Pair 14



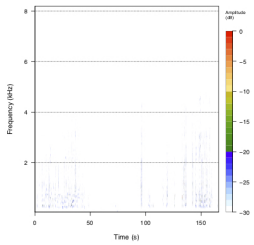

(l) Spectrogram of Listener, Pair 14



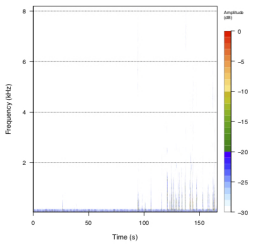

(m) Waveform of Speaker, Pair 22




(n) Waveform of Listener, Pair 22




(o) Spectrogram of Speaker, Pair 22




(p) Spectrogram of Listener, Pair 22

*Figure* 3.1: Waveforms and spectrograms of the speaker and listener from the third, sixth, fourteenth, and twenty-second dyads, or pairs of participants. The plots depict the waveforms after the DC offset has been removed, and before the bandpass filter has been applied. Note that the audio clips for pair 6 and the speaker clip for pair 14 are quieter than the other audio clips shown.

## 3.2 Windowed Cross-Correlations

### 3.2.1 Overall Results

Descriptive statistics of peak lags from the four types of WCC analyses for all original pairs can be found in Table 3.1 and Figure 3.2 depicts histograms of peak lags from the same WCC analyses. Overall, the average absolute maximum peak lag for the original pair WCC analyses between speaker and listener facial expressions was about 2.1 seconds ($M_{Max}$ = 1.77, $M_{Min}$ = -1.76). Also, the absolute average absolute peak lag for the original pair WCC analyses between a speaker's speech patterns and a listener's facial expressions was about 1.6 seconds ($M_{Max}$ = 1.33, $M_{Min}$ = -1.41), and the average absolute peak lag for the original pair WCC analyses between a speaker's facial expressions and a listener's speech patterns was about 1.7 seconds ($M_{Max}$ = 1.31, $M_{Min}$ = -1.44). Finally, the average absolute maximum peak lag for the original pair WCC analyses between speaker and listener speech patterns was about 1.7 seconds ($M_{Max}$ = 1.45, $M_{Min}$ = -1.39).

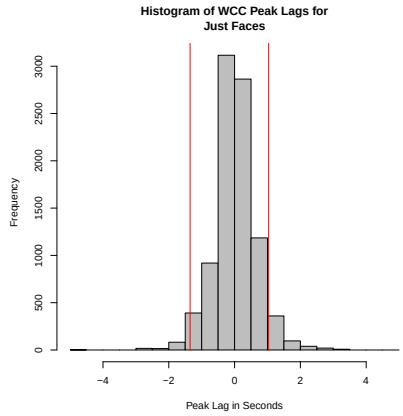
Descriptive statistics of peak lags from the four types of WCC analyses for all surrogate pairs can be found in Table 3.2 and Figure 3.3 depicts histograms of peak lags from the same WCC analyses. In comparison to the original pairs, the average absolute maximum peak lag for the surrogate WCC analyses between speaker and listener facial expressions was also about 2.1 seconds ($M_{Max}$ = 1.71, $M_{Min}$ = -1.73). In addition, the absolute average absolute peak lag for the surrogate WCC analyses between a speaker's speech patterns and a listener's facial expressions was about 1.6 seconds ($M_{Max}$ = 1.38, $M_{Min}$ = -1.37), and the average absolute peak lag for the surrogate WCC analyses between a speaker's facial expressions and a listener's speech patterns was about 1.7 seconds ($M_{Max}$ = 1.44, $M_{Min}$ = -1.43). Finally, the average absolute maximum peak lag for the surrogate WCC analyses between speaker and listener speech patterns was about 1.4 seconds ($M_{Max}$ = 1.14, $M_{Min}$ = -1.17). From these results, there does not seem to be much, if any difference, between the distributions of peak lags between the original and surrogate pairs.

| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Just Faces | 0.02 | 0.64 | 0.02 | -4.71 | 3.46 |
| Speech Sp, Face Lis | 0.01 | 0.54 | 0.00 | -3.41 | 2.50 |
| Face Sp, Speech Lis | -0.02 | 0.58 | -0.02 | -4.39 | 2.89 |
| Just Speech | 0.03 | 0.58 | 0.00 | -3.62 | 4.62 |

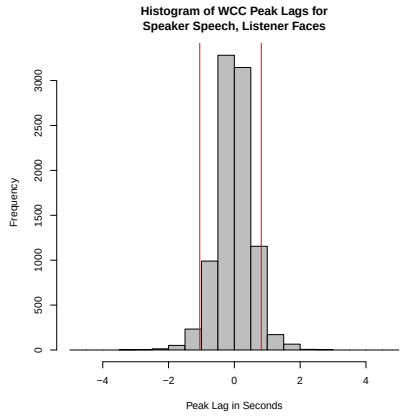
Table 3.1: Descriptive statistics for peak lags from the WCC for the original pairs.

| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Just Faces, Surrogate | -0.00 | 0.65 | -0.05 | -4.92 | 4.61 |
| Speech Sp, Face Lis, Surrogate | -0.00 | 0.51 | -0.02 | -4.64 | 4.39 |
| Face Sp, Speech Lis, Surrogate | -0.01 | 0.54 | -0.03 | -4.66 | 4.81 |
| Just Speech, Surrogate | -0.01 | 0.43 | -0.02 | -4.36 | 3.66 |

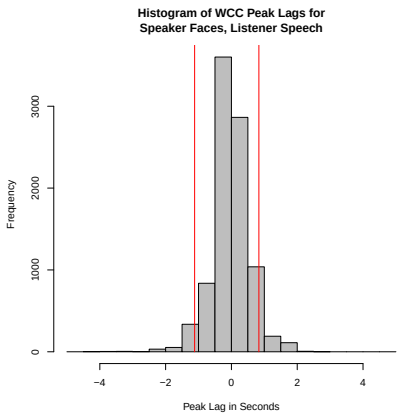
Table 3.2: Descriptive statistics for peak lags from the WCC for the surrogate pairs.



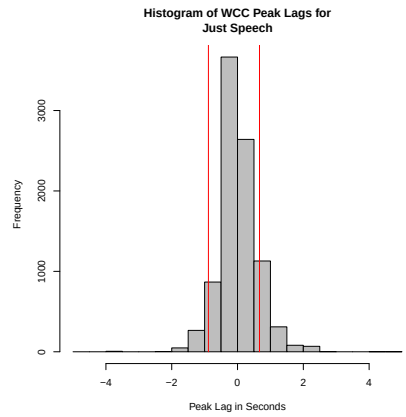

(a) Speaker Face, Listener Face WCC Peak Lag Histogram




(b) Speaker Speech, Listener Face WCC Peak Lag Histogram

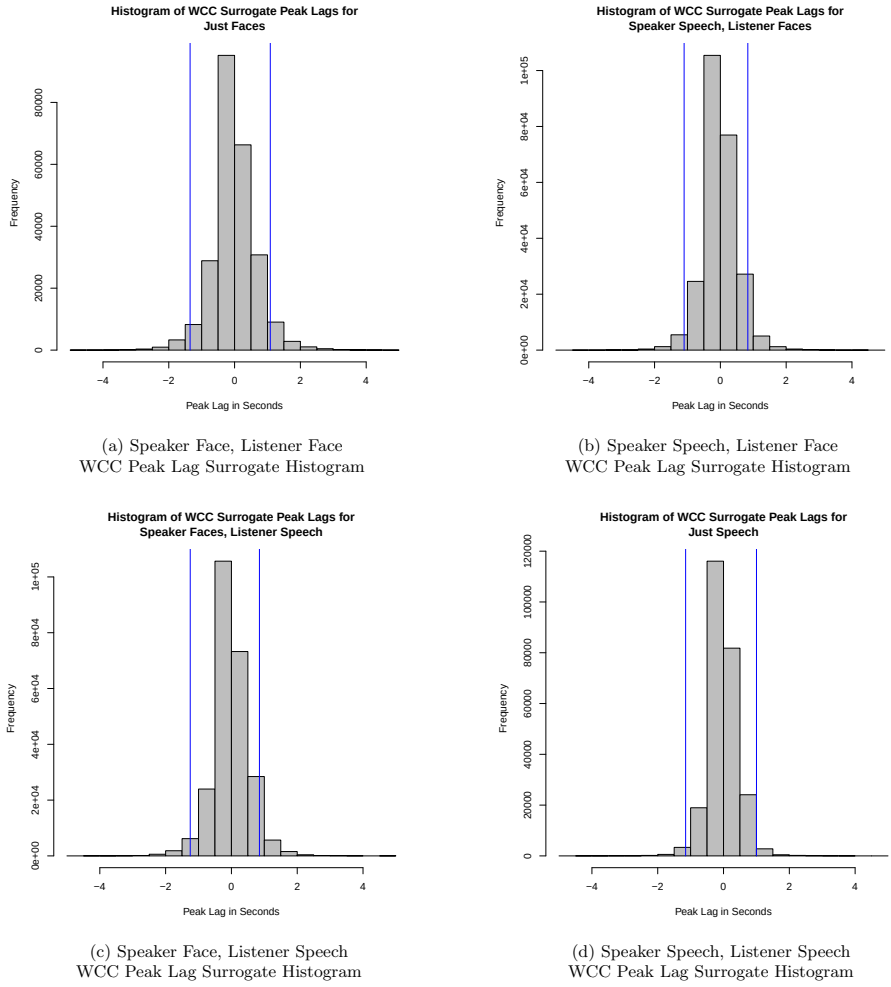



(c) Speaker Face, Listener Speech WCC Peak Lag Histogram




(d) Speaker Speech, Listener Speech WCC Peak Lag Histogram

*Figure* 3.2: Histograms of peak lags from the WCC for the original pairs. The red bars represent where 95% of the peak lags for the surrogate pairs are.




(a) Speaker Face, Listener Face WCC Peak Lag Surrogate Histogram

(b) Speaker Speech, Listener Face WCC Peak Lag Surrogate Histogram

(c) Speaker Face, Listener Speech WCC Peak Lag Surrogate Histogram

(d) Speaker Speech, Listener Speech WCC Peak Lag Surrogate Histogram

*Figure* 3.3: Histograms of peak lags from the WCC for the surrogate pairs. The blue bars represent where 95% of the peak lags for the original pairs are.

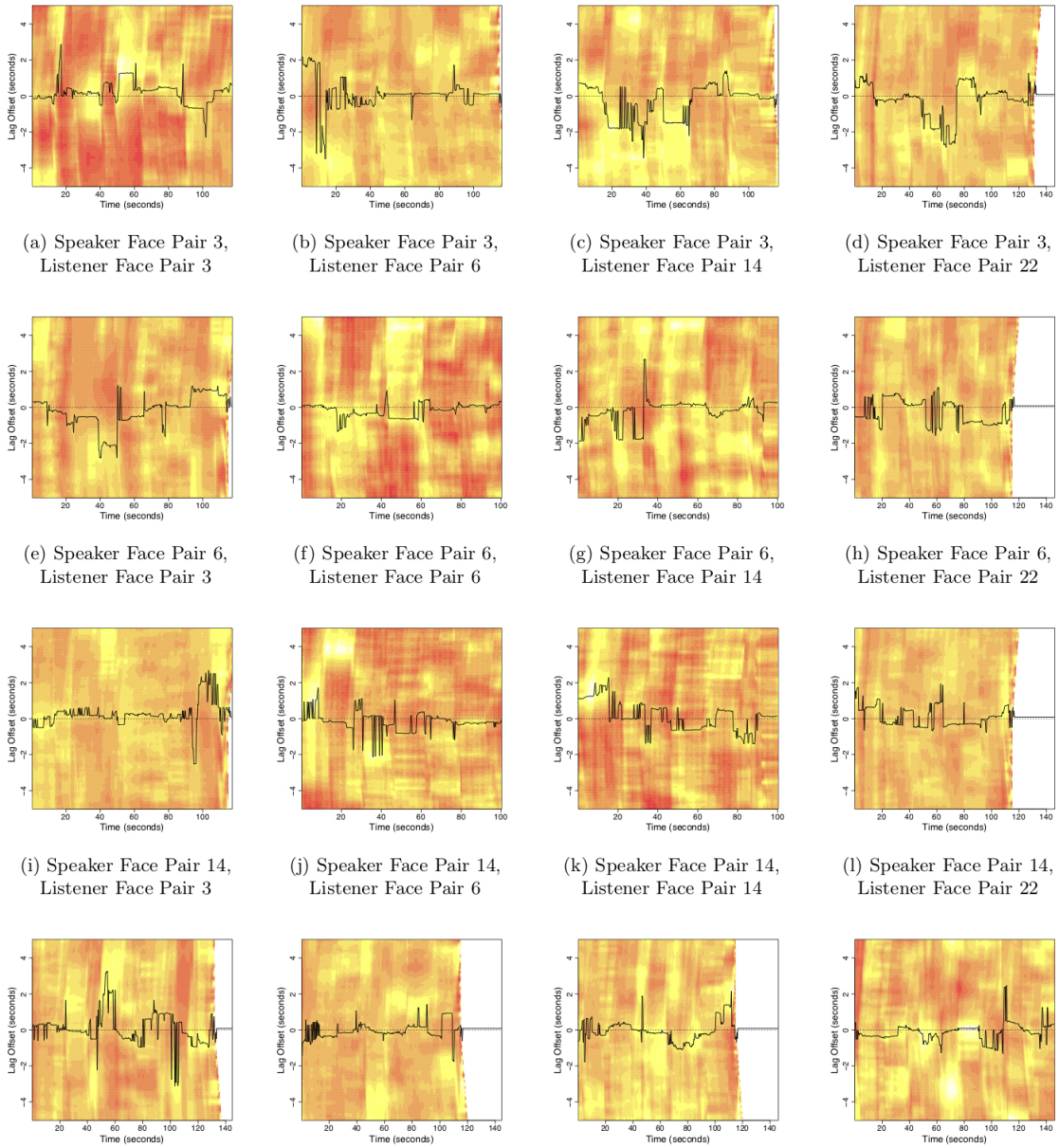
### 3.2.2 Per-Pair Results

Figures 3.4, 3.5, 3.6, and 3.7 show the windowed cross-correlation heatmap plots for pairs 3, 6, 14, and 22 and their surrogates between both participants' faces, between the designated speaker's speech and the designated listener's face, between the designated speaker's face and the designated listener's speech, and between both participants' speech, respectively. As noted before, a strong negative correlation is represented by red coloration in the heatmaps, while a strong positive correlation is represented by yellow.

When looking at the plots that compare the faces of the two participants, there ap-

pear to be periods of stability of the peak lag and correlations throughout the conversations. Although the number and length of these stable periods differ between the original pairs and the surrogates, though, there is not any strong evidence towards whether the original pairs have more or less symmetry formation and breaking than the surrogates. Similarly, there does not seem to be a significant difference in the number of sign changes in the lags between the original pairs and the surrogates.

The results found in the Face-Face analyses are mostly similar to the analyses that include the speech of at least one other participant. When looking at these plots, there appear to be periods of stability of the peak lag and correlations throughout the conversations, and in most cases the presence of these periods of stability does not differ greatly between the original pairs and the surrogates. However, in the plots of these three types of analyses, there are clear sections of time where the patterns of coloration in the plots change around the same time a participant starts and stops talking. For example, in the plots that correspond to the original third pair (Subfigure (a)), there is a clear band of coloration that differs from the rest of the heatmap. This band corresponds to the time when the listener talks and the speaker briefly stops talking. Similarly, in the plots that correspond to the original twenty-second pair (Subfigure (p)), there is a clear change in coloration between the time when no one is speaking, and when both participants start speaking more. For the plots that correspond to the analyses that compare the speech of the designated speaker and the speech of the listener, these changes in correlations are most prominent. In general, this at least provides evidence of turn-taking that is dependent on the presence or absence of speech in a conversation.




(a) Speaker Face Pair 3, Listener Face Pair 3

(b) Speaker Face Pair 3, Listener Face Pair 6

(c) Speaker Face Pair 3, Listener Face Pair 14

(d) Speaker Face Pair 3, Listener Face Pair 22

(e) Speaker Face Pair 6, Listener Face Pair 3

(f) Speaker Face Pair 6, Listener Face Pair 6

(g) Speaker Face Pair 6, Listener Face Pair 14

(h) Speaker Face Pair 6, Listener Face Pair 22

(i) Speaker Face Pair 14, Listener Face Pair 3

(j) Speaker Face Pair 14, Listener Face Pair 6

(k) Speaker Face Pair 14, Listener Face Pair 14

(l) Speaker Face Pair 14, Listener Face Pair 22

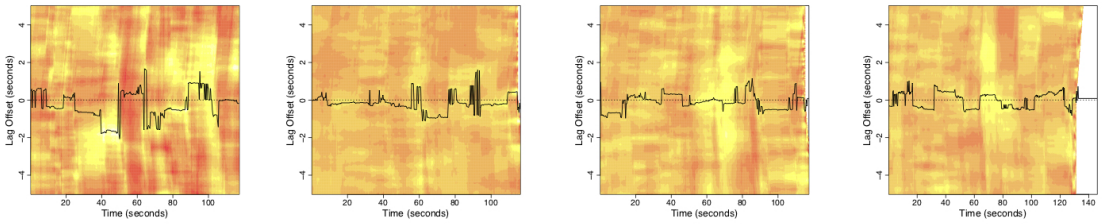
(m) Speaker Face Pair 22, Listener Face Pair 3

(n) Speaker Face Pair 22, Listener Face Pair 6

(o) Speaker Face Pair 22, Listener Face Pair 14

(p) Speaker Face Pair 22, Listener Face Pair 22

*Figure* 3.4: Heatmap plots of the WCC between the designated speaker's face and the designated listener's face for four representative pairs and their respective surrogates.

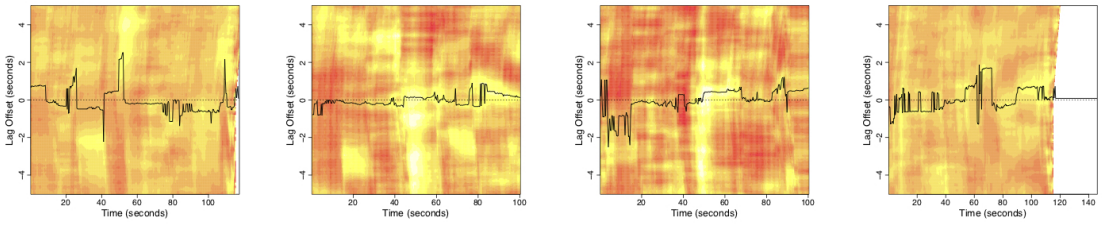



(a) Speaker Speech Pair 3, Listener Face Pair 3

(b) Speaker Speech Pair 3, Listener Face Pair 6

(c) Speaker Speech Pair 3, Listener Face Pair 14

(d) Speaker Speech Pair 3, Listener Face Pair 22

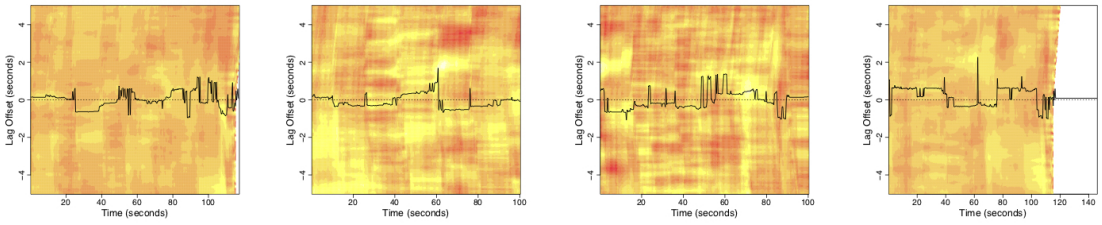



(e) Speaker Speech Pair 6, Listener Face Pair 3

(f) Speaker Speech Pair 6, Listener Face Pair 6

(g) Speaker Speech Pair 6, Listener Face Pair 14

(h) Speaker Speech Pair 6, Listener Face Pair 22

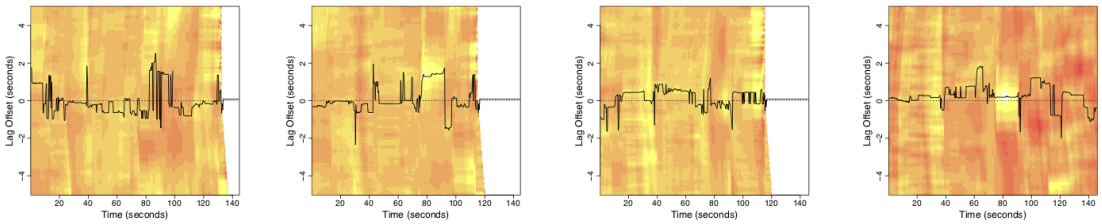



(i) Speaker Speech Pair 14, Listener Face Pair 3

(j) Speaker Speech Pair 14, Listener Face Pair 6

(k) Speaker Speech Pair 14, Listener Face Pair 14

(l) Speaker Speech Pair 14, Listener Face Pair 22



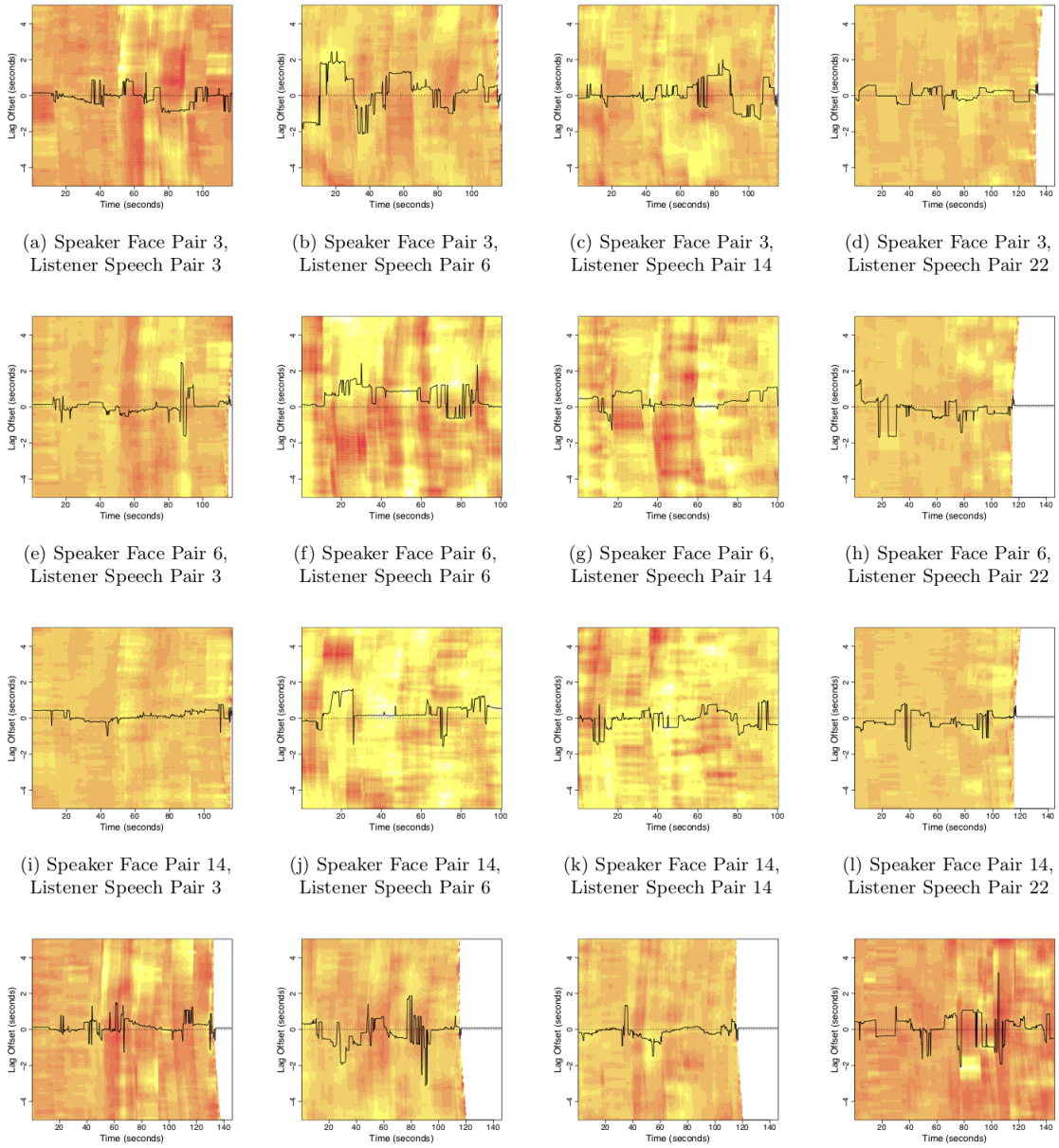

(m) Speaker Speech Pair 22, Listener Face Pair 3

(n) Speaker Speech Pair 22, Listener Face Pair 6

(o) Speaker Speech Pair 22, Listener Face Pair 14

(p) Speaker Speech Pair 22, Listener Face Pair 22

*Figure* 3.5: Heatmap plots of the WCC between the designated speaker's speech and the designated listener's face for four representative pairs and their respective surrogates.




(a) Speaker Face Pair 3, Listener Speech Pair 3

(b) Speaker Face Pair 3, Listener Speech Pair 6

(c) Speaker Face Pair 3, Listener Speech Pair 14

(d) Speaker Face Pair 3, Listener Speech Pair 22

(e) Speaker Face Pair 6, Listener Speech Pair 3

(f) Speaker Face Pair 6, Listener Speech Pair 6

(g) Speaker Face Pair 6, Listener Speech Pair 14

(h) Speaker Face Pair 6, Listener Speech Pair 22

(i) Speaker Face Pair 14, Listener Speech Pair 3

(j) Speaker Face Pair 14, Listener Speech Pair 6

(k) Speaker Face Pair 14, Listener Speech Pair 14

(l) Speaker Face Pair 14, Listener Speech Pair 22

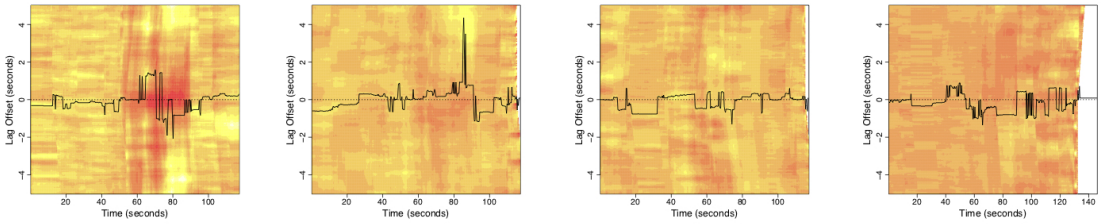
(m) Speaker Face Pair 22, Listener Speech Pair 3

(n) Speaker Face Pair 22, Listener Speech Pair 6

(o) Speaker Face Pair 22, Listener Speech Pair 14

(p) Speaker Face Pair 22, Listener Speech Pair 22

*Figure* 3.6: Heatmap plots of the WCC between the designated speaker's face and the designated listener's speech for four representative pairs and their respective surrogates.

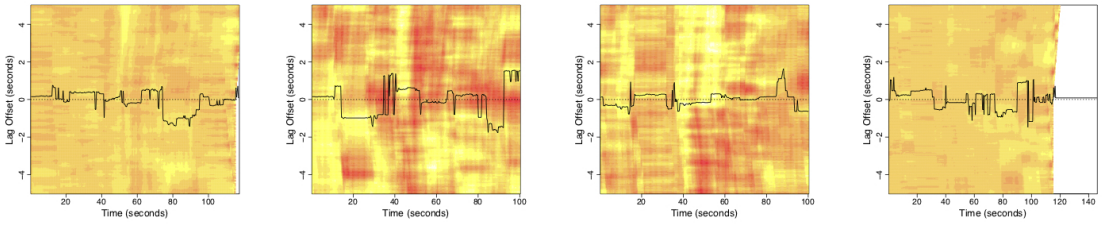



(a) Speaker Speech Pair 3, Listener Speech Pair 3

(b) Speaker Speech Pair 3, Listener Speech Pair 6

(c) Speaker Speech Pair 3, Listener Speech Pair 14

(d) Speaker Speech Pair 3, Listener Speech Pair 22

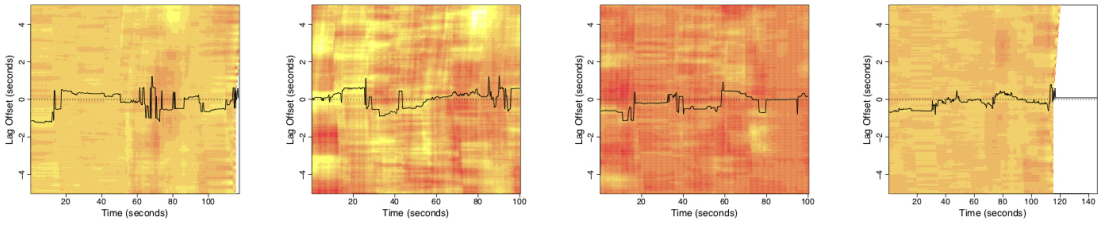



(e) Speaker Speech Pair 6, Listener Speech Pair 3

(f) Speaker Speech Pair 6, Listener Speech Pair 6

(g) Speaker Speech Pair 6, Listener Speech Pair 14

(h) Speaker Speech Pair 6, Listener Speech Pair 22

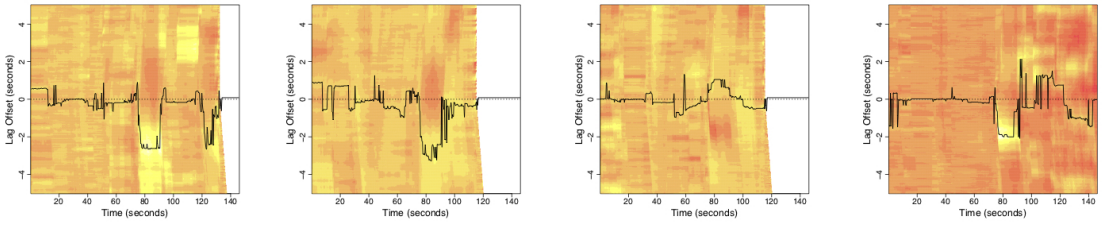



(i) Speaker Speech Pair 14, Listener Speech Pair 3

(j) Speaker Speech Pair 14, Listener Speech Pair 6

(k) Speaker Speech Pair 14, Listener Speech Pair 14

(l) Speaker Speech Pair 14, Listener Speech Pair 22




(m) Speaker Speech Pair 22, Listener Speech Pair 3

(n) Speaker Speech Pair 22, Listener Speech Pair 6

(o) Speaker Speech Pair 22, Listener Speech Pair 14

(p) Speaker Speech Pair 22, Listener Speech Pair 22

*Figure* 3.7: Heatmap plots of the WCC between the designated speaker's speech and the designated listener's speech for four representative pairs and their respective surrogates.

## 3.3 Canonical Correlation Analyses

### 3.3.1 Checking of correlations

Table 3.3 and Figure 3.8 provide descriptive statistics and histograms, respectively, for the correlations after performing the four types of analyses on the original pairs when using the full matrices. Because the correlations for the face-face analyses are so high – all correlations for the face-face original pair analyses are greater than .9 – the face-face CCA were run again with the first 12 components from the left singular matrices that were constructed from performing SVD decomposition on the point coordinate matrices. Almost all of the correlations from the other three analyses were lower than the .9 threshold ($Percentage_{speakerspeech,listenerface<.9} = 99.79\%$, $Percentage_{speakerface,listenerspeech<.9} = 99.59\%$, $Percentage_{speakerspeech,listenerspeech<.9} = 99.99\%$), and so decomposing the STFT matrices to get the respective left singular matrices for the speech patterns was not required. However, to maintain consistency, both the full matrices and the left singular matrices for both the point coordinate and STFT matrices were used for all four types of analyses for the original pairs. For the surrogates, left singular matrices were used for the face-face analyses, and the full matrices were used for the other three analyses.

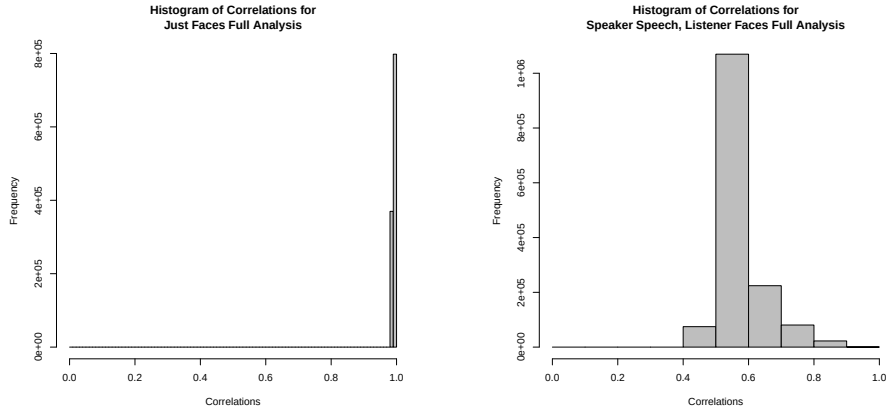
Table 3.4 and Figure 3.9 provide the new descriptive statistics and histograms, respectively, for the correlations after doing the same analyses using the left singular matrix analyses. As expected, the correlations for these analyses tend to be lower than their corresponding analyses using the full matrices ($Percentage_{speakerface,listenerface<.9} = 97.68\%$, $Percentage_{speakerspeech,listenerface<.9} > 99.99\%$, $Percentage_{speakerface,listenerspeech<.9} > 99.99\%$, $Percentage_{speakerspeech,listenerspeech<.9} > 99.99\%$).

| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Full, Just Faces, Cor | 0.99 | 0.00 | 0.99 | 0.98 | 1.00 |
| Full, Speech Sp, Face Lis, Cor | 0.62 | 0.10 | 0.60 | 0.42 | 0.97 |
| Full, Face Sp, Speech Lis, Cor | 0.62 | 0.09 | 0.59 | 0.44 | 0.97 |
| Full, Just Speech, Cor | 0.55 | 0.05 | 0.54 | 0.46 | 0.99 |

Table 3.3: Descriptive statistics for correlations from the CCA for the analyses that used the full matrices.

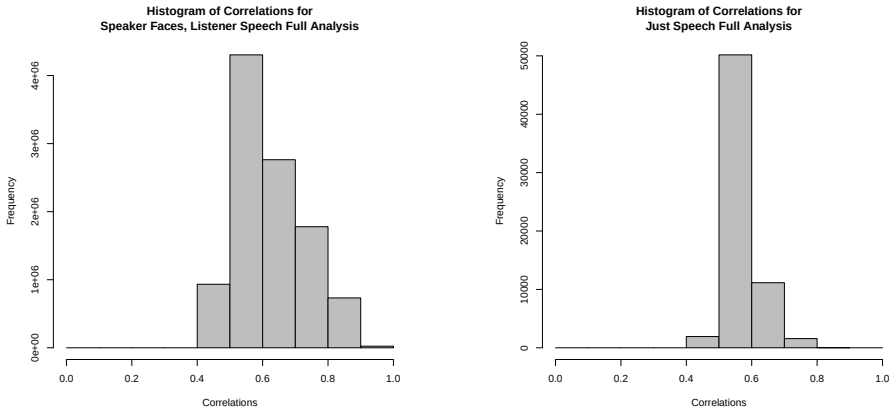
| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Left Singular, Just Faces, Cor | 0.53 | 0.21 | 0.52 | 0.09 | 0.99 |
| Left Singular, Speech Sp, Face Lis, Cor | 0.30 | 0.15 | 0.28 | 0.05 | 0.92 |
| Left Singular, Face Sp, Speech Lis, Cor | 0.29 | 0.16 | 0.25 | 0.05 | 0.96 |
| Left Singular, Just Speech, Cor | 0.23 | 0.09 | 0.21 | 0.05 | 0.98 |

Table 3.4: Descriptive statistics for correlations from the CCA for the analyses that used the left singular matrices.




(a) Speaker Face, Listener Face
CCA Correlation Histogram, Full Matrices

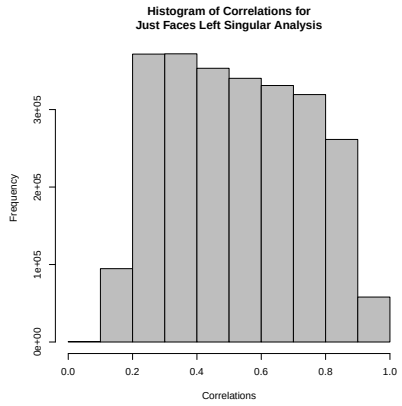
(b) Speaker Speech, Listener Face
CCA Correlation Histogram, Full Matrices



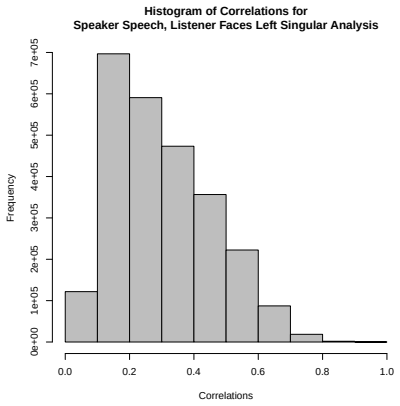

(c) Speaker Face, Listener Speech
CCA Correlation Histogram, Full Matrices

(d) Speaker Speech, Listener Speech
CCA Correlation Histogram, Full Matrices

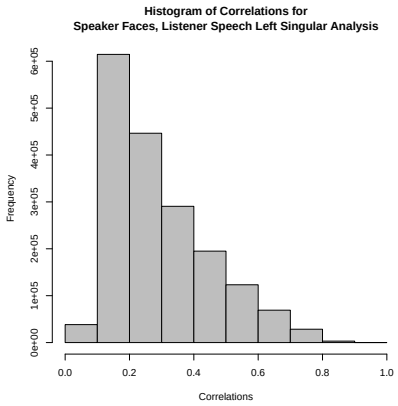
*Figure* 3.8: Histograms of correlations from the CCA when using the full matrices.



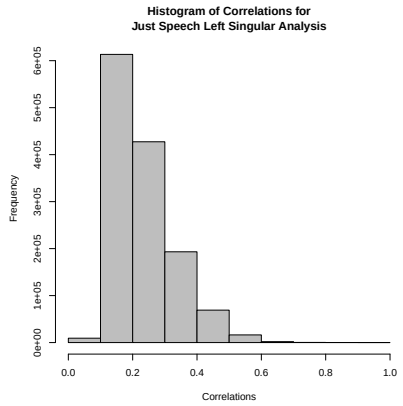

(a) Speaker Face, Listener Face
CCA Correlation Histogram, Left Singular Matrices




(b) Speaker Speech, Listener Face
CCA Correlation Histogram, Left Singular Matrices




(c) Speaker Face, Listener Speech
CCA Correlation Histogram, Left Singular Matrices




(d) Speaker Speech, Listener Speech
CCA Correlation Histogram, Left Singular Matrices

*Figure* 3.9: Histograms of correlations from the CCA when using the left singular matrices.

### 3.3.2 Overall Results

#### Full Analyses

Descriptive statistics of significant canonical pairs from the four types of CCA analyses for all original pairs using the full matrices can be found in Table 3.5, and Figure 3.10 depicts histograms of significant canonical pairs from the same CCA analyses. Overall, the ranges for the four analyses vary widely, where the number of significant canonical pairs when comparing the faces of both participants and when comparing the speech patterns of both participants are much lower ($M_{speakerface,listenerface}$ = 4.08, $M_{speakerspeech,listenerspeech}$

= 4.11) than for the analyses when comparing one participant's face with another participant's speech ($M_{speakerspeech,listenerface}$ = 37.14, $M_{speakerface,listenerspeech}$ = 18.93). Notably, the distribution of the number of significant canonical pairs for the analyses between a designated speaker's speech and a designated listener's face appears relatively normal, whereas the distributions for the other three analyses are positively skewed. This implies that the dimensionality of the analyses between a designated speaker's speech and a designated listener's face is higher than the other three analyses, which could imply that the speaker's speech is driving the conversation, and potentially any synchrony occurring between the two participants.

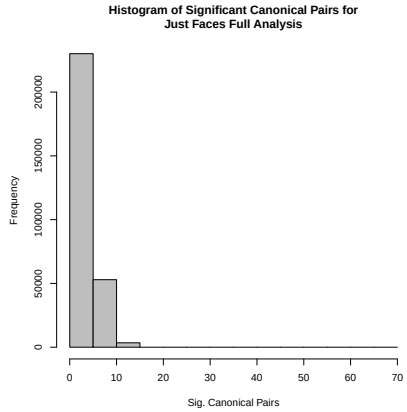
Descriptive statistics of the proportions of variance that are explained by the significant canonical pairs from the four types of CCA analyses for all original pairs using the full matrices can be found in Table 3.6, and Figure 3.11 depicts histograms of the proportions of variance explained from the same CCA analyses. In general, the ranges for the proportion of variance explained for all four analyses are very small, which may imply that any synchrony or turn-taking found may be mostly driven by the number of significant canonical pairs, or the dimensionality of the analyses, rather than the corresponding proportion of explained variance.

| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Full, Just Faces, Sig. Pairs | 4.08 | 2.06 | 4 | 0 | 15 |
| Full, Speech Sp, Face Lis, Sig. Pairs | 37.14 | 12.73 | 38 | 4 | 66 |
| Full, Face Sp, Speech Lis, Sig. Pairs | 18.93 | 11.96 | 15 | 3 | 61 |
| Full, Just Speech, Sig. Pairs | 4.11 | 3.18 | 3 | 0 | 23 |

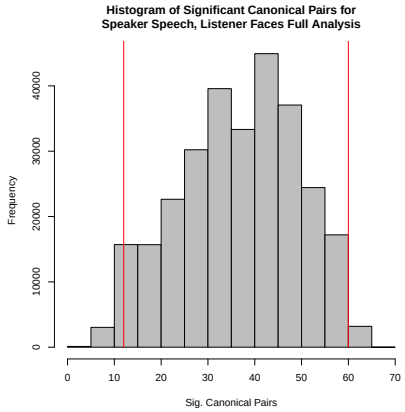
Table 3.5: Descriptive statistics for the number of significant canonical pairs from the CCA when using the full matrices.

| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Full, Just Faces, Prop | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 |
| Full, Speech Sp, Face Lis, Prop | 0.01 | 0.00 | 0.01 | 0.01 | 0.02 |
| Full, Face Sp, Speech Lis, Prop | 0.01 | 0.00 | 0.01 | 0.01 | 0.02 |
| Full, Just Speech, Prop | 0.02 | 0.00 | 0.02 | 0.01 | 0.03 |

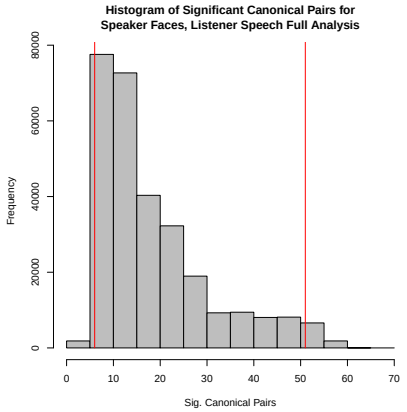
Table 3.6: Descriptive statistics for the proportion of variance explained by the significant canonical pairs from the CCA when using the full matrices.



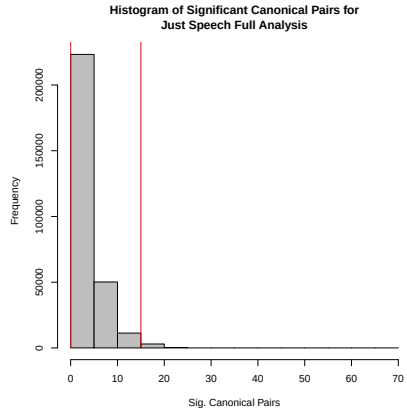

(a) Speaker Face, Listener Face
CCA Sig. Pairs Histogram, Full Matrices




(b) Speaker Speech, Listener Face
CCA Sig. Pairs Histogram, Full Matrices

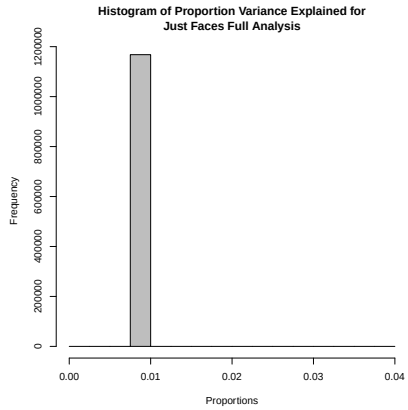



(c) Speaker Face, Listener Speech
CCA Sig. Pairs Histogram, Full Matrices

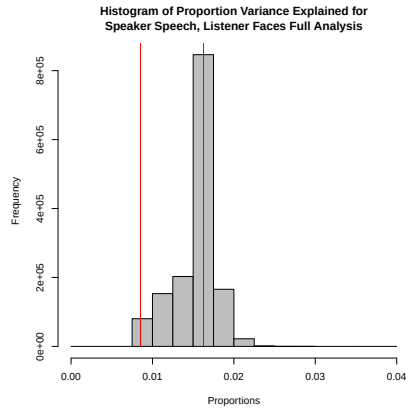



(d) Speaker Speech, Listener Speech
CCA Sig. Pairs Histogram, Full Matrices

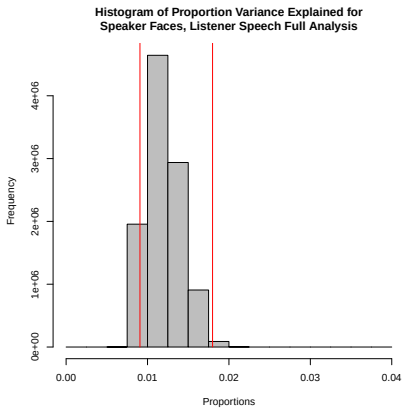
*Figure* 3.10: Histograms of the number of significant canonical pairs from the CCA when using the full matrices. The red bars represent where 95% of the peak lags for the surrogate pairs are. Plots with no bars do not have a corresponding set of surrogate analyses.



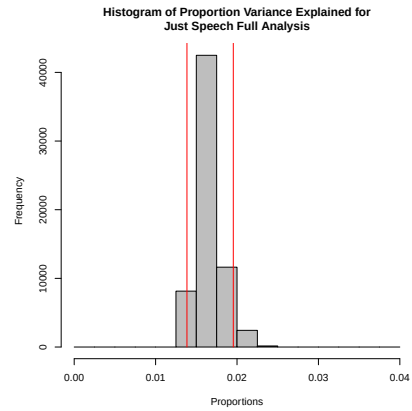

(a) Speaker Face, Listener Face
CCA Proportion Histogram, Full Matrices




(b) Speaker Speech, Listener Face
CCA Proportion Histogram, Full Matrices




(c) Speaker Face, Listener Speech
CCA Proportion Histogram, Full Matrices




(d) Speaker Speech, Listener Speech
CCA Proportion Histogram, Full Matrices

*Figure* 3.11: Histograms of the proportion of variance explained from the CCA when using the full matrices. The red bars represent where 95% of the peak lags for the surrogate pairs are. Plots with no bars do not have a corresponding set of surrogate analyses.

### Left Singular Analyses

Descriptive statistics of significant canonical pairs from the four types of CCA analyses for all original pairs using the left singular matrices can be found in Table 3.7, and Figure 3.12 depicts histograms of significant canonical pairs from the same CCA analyses. In all four analyses, due to only using the first twelve columns of the left singular matrices, the maximum number of significant canonical pairs is also 12. When comparing distributions, the analyses that compare the designated listener's face to either the face or speech of the designated speaker are both negatively skewed, while the analyses that compare the

designated speaker's face to either the face or speech of the listener are more normally distributed (designated speaker's face) or positively skewed (designated speaker's speech). This again provides evidence towards the speaker driving the conversation, and potentially any synchrony occurring between the two participants.

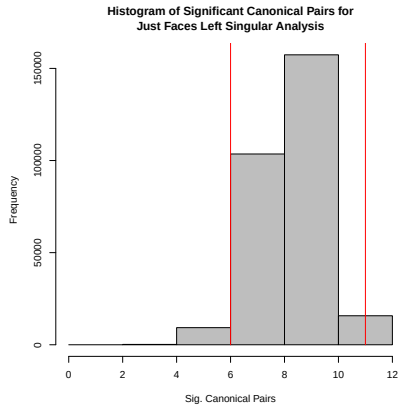
Descriptive statistics of the proportions of variance that are explained by the significant canonical pairs from the four types of CCA analyses for all original pairs using the left singular matrices can be found in Table 3.8, and Figure 3.13 depicts histograms of the proportions of variance explained from the same CCA analyses. From looking at the proportion of variance explained by the significant canonical pairs in the left singular analyses, these analyses tend to have higher proportions of variance than the corresponding full analyses. This is because there is an upper limit of significant canonical pairs due to only using the first twelve columns of these matrices, and so the same amount of variance explained for all canonical pairs in the full analyses is being sectioned into fewer, larger canonical pairs in the left singular analyses. Also, other than the analysis that compares the participants' faces, which has a lower maximum and a more normally distributed distribution, the proportions tend to have similar ranges and positively skewed distributions. These similarities provide more evidence that any synchrony or turn-taking found may be mostly driven by the number of significant canonical pairs, or the dimensionality of the analyses, rather than the corresponding proportion of explained variance.

| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Left Singular, Just Faces, Sig. Pairs | 8.75 | 1.19 | 9 | 3 | 12 |
| Left Singular, Speech Sp, Face Lis, Sig. Pairs | 8.95 | 1.60 | 9 | 1 | 12 |
| Left Singular, Face Sp, Speech Lis, Sig. Pairs | 6.30 | 2.38 | 6 | 0 | 12 |
| Left Singular, Just Speech, Sig. Pairs | 4.62 | 2.09 | 4 | 0 | 12 |

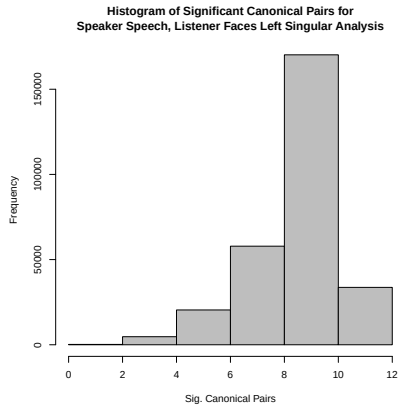
Table 3.7: Descriptive statistics for the number of significant canonical pairs from the CCA when using the left singular matrices.

| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Left Singular, Just Faces, Prop | 0.11 | 0.04 | 0.11 | 0.01 | 0.28 |
| Left Singular, Speech Sp, Face Lis, Prop | 0.11 | 0.05 | 0.10 | 0.01 | 0.47 |
| Left Singular, Face Sp, Speech Lis, Prop | 0.13 | 0.07 | 0.11 | 0.01 | 0.52 |
| Left Singular, Just Speech, Prop | 0.14 | 0.06 | 0.13 | 0.01 | 0.58 |

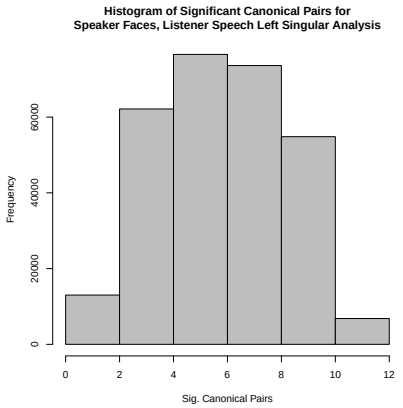
Table 3.8: Descriptive statistics for the proportion of variance explained by the significant canonical pairs from the CCA when using the left singular matrices.



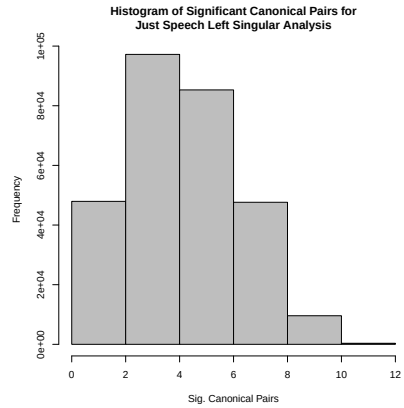

(a) Speaker Face, Listener Face
CCA Sig. Pairs Histogram, Left Singular Matrices




(b) Speaker Speech, Listener Face
CCA Sig. Pairs Histogram, Left Singular Matrices

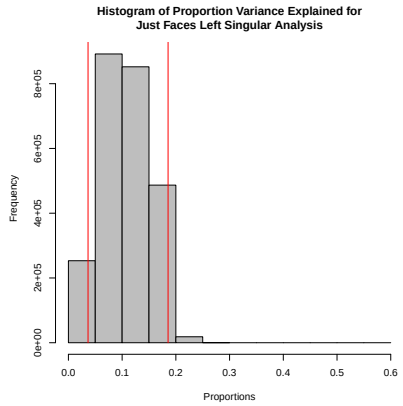



(c) Speaker Face, Listener Speech
CCA Sig. Pairs Histogram, Left Singular Matrices

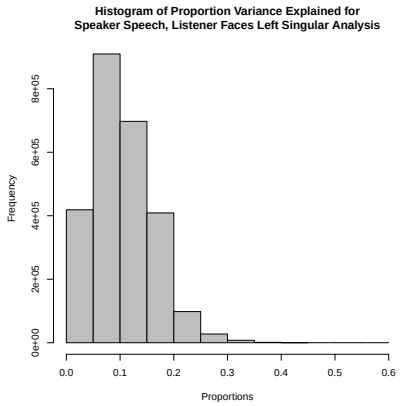



(d) Speaker Speech, Listener Speech
CCA Sig. Pairs Histogram, Left Singular Matrices

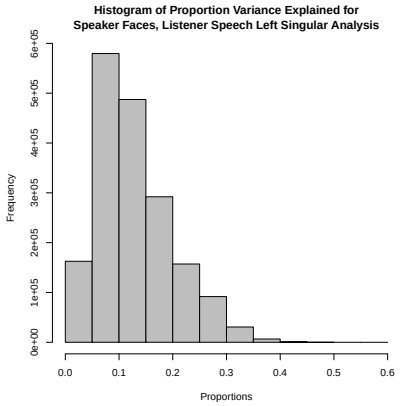
*Figure* 3.12: Histograms of the number of significant canonical pairs from the CCA when using the left singular matrices. The red bars represent where 95% of the peak lags for the surrogate pairs are. Plots with no bars do not have a corresponding set of surrogate analyses.



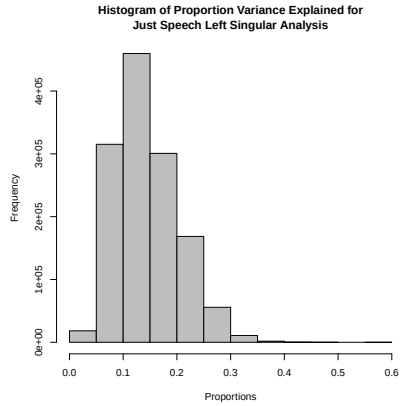

(a) Speaker Face, Listener Face
CCA Proportion Histogram, Left Singular Matrices




(b) Speaker Speech, Listener Face
CCA Proportion Histogram, Left Singular Matrices




(c) Speaker Face, Listener Speech
CCA Proportion Histogram, Left Singular Matrices




(d) Speaker Speech, Listener Speech
CCA Proportion Histogram, Left Singular Matrices

*Figure* 3.13: Histograms of the proportion of variance explained from the CCA when using the left singular matrices. The red bars represent where 95% of the peak lags for the surrogate pairs are. Plots with no bars do not have a corresponding set of surrogate analyses.

## Surrogate Analyses

Descriptive statistics of significant canonical pairs from the four types of CCA analyses for all surrogate pairs can be found in Table 3.9, and Figure 3.14 depicts histograms of significant canonical pairs from the same CCA analyses. Also, descriptive statistics of the proportions of variance that are explained by the significant canonical pairs from the four types of CCA analyses for all surrogates can be found in Table 3.10, and Figure 3.15 depicts histograms of the proportions of variance explained from the same CCA analyses. As noted previously, the surrogates when comparing the faces of the participants in the

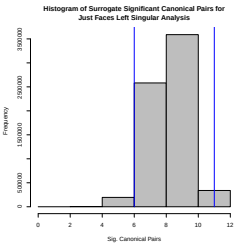
conversation were analyzed using the left singular matrices, while the other three surrogate types of analyses used the full matrices. For all analyses, both the ranges and distributions of the surrogates for the number of significant canonical pairs as well as the corresponding proportion of variance explained are fairly close to the respective distributions from the original pairs. This therefore does not provide sufficient evidence towards the presence of synchrony being dependent on the conversations themselves when aggregating across all of the pairs, but rather on the general properties of facial expression movement and speech pattern variation.

| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Left Sing, Just Faces, Sig. Pairs | 8.71 | 1.17 | 9 | 2 | 12 |
| Full, Speech Sp, Face Lis, Sig. Pairs | 37.85 | 12.61 | 39 | 3 | 67 |
| Full, Face Sp, Speech Lis, Sig. Pairs | 18.75 | 12.28 | 14 | 3 | 62 |
| Full, Just Speech, Sig. Pairs | 4.45 | 3.91 | 3 | 0 | 28 |

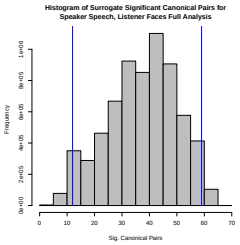
Table 3.9: Descriptive statistics for the number of significant canonical pairs from the CCA for all analyses on the surrogate pairs.

| | mean | sd | median | min | max |
|---|---|---|---|---|---|
| Left Sing, Just Faces, Prop | 0.11 | 0.04 | 0.11 | 0.01 | 0.37 |
| Full, Speech Sp, Face Lis, Prop | 0.01 | 0.00 | 0.01 | 0.01 | 0.02 |
| Full, Face Sp, Speech Lis, Prop | 0.01 | 0.00 | 0.01 | 0.01 | 0.02 |
| Full, Just Speech, Prop | 0.02 | 0.00 | 0.02 | 0.01 | 0.03 |

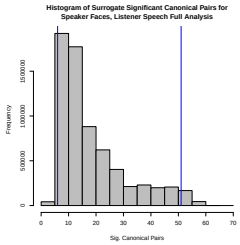
Table 3.10: Descriptive statistics for the proportion of variance explained by the significant canonical pairs from the CCA for all analyses on the surrogate pairs.



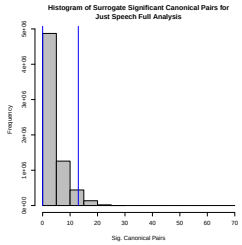

(a) Speaker Face, Listener Face CCA Sig. Pairs Surrogate Histogram, Left Singular Matrices




(b) Speaker Speech, Listener Face CCA Sig. Pairs Surrogate Histogram, Full Matrices

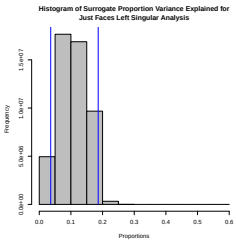



(c) Speaker Face, Listener Speech CCA Sig. Pairs Surrogate Histogram, Full Matrices

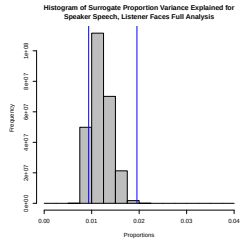



(d) Speaker Speech, Listener Speech CCA Sig. Pairs Surrogate Histogram, Full Matrices

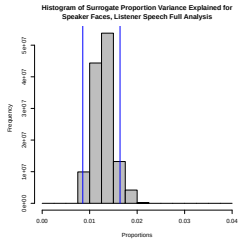
*Figure* 3.14: Histograms of the number of significant canonical pairs from the CCA for all analyses on the surrogate pairs. The blue bars represent where 95% of the peak lags for the original pairs are.



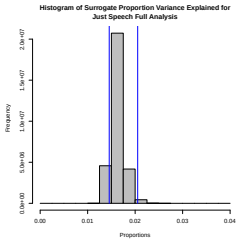

(a) Speaker Face, Listener Face CCA Proportion Surrogate Histogram, Left Singular Matrices




(b) Speaker Speech, Listener Face CCA Proportion Surrogate Histogram, Full Matrices




(c) Speaker Face, Listener Speech CCA Proportion Surrogate Histogram, Full Matrices




(d) Speaker Speech, Listener Speech CCA Proportion Surrogate Histogram, Full Matrices

*Figure* 3.15: Histograms of the proportion of variance explained from the CCA for all analyses on the surrogate pairs. The blue bars represent where 95% of the peak lags for the original pairs are.

### 3.3.3 Per-Pair Results

#### Full Analyses

Plots of the significant canonical pairs across both the original and surrogate conversations when using the full data matrices can be seen in Figures A.1, A.2, A.3, and A.4 in Appendix A. These plots specifically show how the number of significant canonical pairs change between windows and across lags for pairs 3, 6, 14, and 22 and their surrogates between both participants' faces, between the designated speaker's speech and the designated listener's face, between the designated speaker's face and the designated listener's speech, and between both participants' speech, respectively. In these plots, each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.

Overall, when comparing the plots that represent the CCA between the faces of the two participants, the number of significant canonical pairs remains relatively constant across time. However, there are slight changes in these numbers that do seem to be affected by what is happening in the conversations. For example in pair 3, when the designated speaker stops talking and the designated listener starts, there is a noticeable decrease in the number of significant canonical pairs; the average number of significant canonical pairs then increases back to its previous amount once the speaking roles switch back.

When comparing speech from the designated speaker to the designated listener's face, the curvature of the plots tend to match up with the amount of talking that the designated speaker is doing. Similarly, the curvature of the plots that depict the relationship between the designated speaker's face and the designated listener's speech are mainly dependent on how much talking the designated listener is doing. In both types of analyses, as the participant whose speech patterns are being analyzed talks more, the number of significant canonical pairs increases, and as that participant talks less, that number tends to decrease. Also, in many cases for at least the analyses between one participant's face and

another participant's speech patterns, whenever the absolute value of the magnitude of the curvature's slope is high, the lags tend to have similar curvatures with each other. Because the curvatures are so dependent on the speech patterns analyzed, there is no significant difference, other than slight differences in the different lag curvatures, between the number of significant canonical pairs between the original pairs and the surrogates when the faces of the original pairs are swapped with surrogate faces. When speech patterns of the original pairs are swapped with surrogate speech patterns, though, the curvatures between the original pairs and the surrogates across all lags noticeably differ.

Finally, when comparing the speech of the designated speaker with the speech of the designated listener, the number of significant canonical pairs appears to be mainly dependent on both participants' speaking roles; namely, when either, or both participants talk. Whenever both participants are talking, the number of significant canonical pairs increases, and when only one or neither participants talk, this number decreases. Possibly due to this, the average number of significant canonical pairs tends to be lower than for the other three analyses. When comparing the analyses on the original pairs with those on the surrogates, whether they differ seem to be due to differences in the amount of talking between the original participants and the surrogate participants. These analyses therefore provide more evidence towards the presence of turn-taking, at least with regards to speaking roles and the presence or absence of talking in conversations.

Similar CCA per-pair plots for the total, or cumulative proportion of variance explained by the significant canonical pairs from the full analysis CCA can be found in Figures A.5, A.6, A.7, and A.8 in Appendix A. Figures A.9 through A.12 represent the average proportion of variance explained for the significant canonical pairs, and Figures A.13 through A.16 represent the largest proportion of variance explained for the significant canonical pairs.

In terms of variance explained, the total, or cumulative proportion of variance explained by the significant canonical pairs tend to have a similar curvature across the con-

versation as the number of significant canonical pairs for all analyses. Also, in general, the total variance of proportion explained tended to have more variation across the lags than the corresponding number of significant pairs for the same analyses. Potentially due to the number of significant canonical pairs having similar curvatures to the corresponding cumulative proportion of variance explained, neither the average nor the largest proportion of variance explained by the significant canonical pairs tend to change significantly (i.e., more than 2%) across time. However, within this range, both the average and largest proportion of variance explained do change slightly across time. Overall, the proportion of variance explained therefore does not provide any more or less evidence towards the presence of synchrony or turn-taking found when looking at the results of the significant canonical pairs.

**Left Singular Analyses**

Figures A.17 through A.20, Figures A.21 through A.24, Figures A.25 through A.28, and Figures A.29 through A.32 in Appendix A represent per-pair plots for the significant canonical pairs, the corresponding total proportion of variance explained, the corresponding average proportion of variance explained, and the corresponding largest proportion of variance explained, respectively, for the CCA analyses when using the left-singular matrices.

When looking at the analyses that used the left singular matrices, the number of significant canonical pairs tended to either be constant at 12 significant canonical pairs, or otherwise follow a similar curvature to the corresponding curvatures of significant canonical pairs in the analyses that used the full matrices. The ceiling effect seems to be most prominent in the analyses that compare the designated speaker's speech with the designated listener's face, which may be due to the fact that the designated speaker tended to talk more, and so potentially needed more than 12 components to represent all the variance in the speech patterns. Also, because all of the analyses that compare the designated

speaker's face with the designated listener's face have relatively constant numbers of significant canonical pairs across the conversations, there does not seem to be any strong differences between faces in the original pairs and the surrogates.

Similarly to the number of significant canonical pairs, most of the analyses show a ceiling effect with the total, or cumulative proportion of variance explained by the significant canonical pairs. When this ceiling effect does not occur, the curvature of total proportion of explained variance is similar to the corresponding total proportion of explained variance for the analyses using the full matrices. Also, the average proportion of explained variance and, to a lesser extent, the largest proportion of explained variance appear to have similar but vertically flipped curvatures with the total proportion of variance explained.

# 4. Discussion

## 4.1 Summary of results

After performing the windowed cross-correlations, limited evidence towards synchrony and turn-taking was found in the conversations. In all four types of analyses comparing faces and speech, there were several instances of symmetry formation and breaking between faces and speech in the plots, which demonstrates the presence of synchrony between the participants. When comparing the four analyses with each other, there were clearer instances of turn-taking present in the results when the speaking roles swapped throughout the conversations. While the amount of synchrony did not appear to differ between the original pairs and the surrogates in most instances, there was more variation in correlations (colors) and peak lag in analyses that include speech depending on these speaking roles between all pairs, including between the original pairs and the surrogates, and the amount of turn-taking did differ noticeably between the pairs as well. The differences visually seen between original pairs and surrogates are most prominent in the analyses that compare the designated speaker's speech and the designated listener's speech, implying that similarities and differences, and therefore any evidence of synchrony and turn-taking, in speech patterns is dependent on the conversation.

When looking at the results of the windowed canonical correlation analyses, there is also evidence towards turn-taking in the conversations. In general, the number of significant canonical pairs, or the dimensionality, of the analyses, as well as the proportion of explained variance did not significantly change when faces from original pairs were swapped with surrogate faces, but they did change when speech patterns from the original pairs were swapped. Also, both when looking at the pairs individually and when aggregating across

pairs, the values of and changes in both the dimensionality and proportions of explained variance tended to be higher with analyses that compared the face of one participant with the speech of the other, as opposed to when comparing both faces, which did not have any changes across time, and when comparing both speech patterns, which were highly dependent again on speaking roles. This at least provides evidence towards turn-taking, at least in the form of the swapping of speaking roles.

One possible reason for the differential in the dimensionality between one participant's face and another participant's speech, as opposed to comparing two faces or two speech patterns, could be that the dimensionality of speech patterns is higher in general than the dimensionality of the faces, and the two are more independent of each other than the dimensionality between two faces or two speech patterns. If this is the case, this would suggest that faces and speech are more synchronous with each other than between a face and a speech. Another reason for these results could be that the dimensionality of the faces does not change across time, whereas the dimensionality of the speech changes quite drastically in contrast. The stability of facial dimensionality, then, could be why the faces are contributing less to the Face-Speech and Speech-Face analyses than the speech patterns do. A third reason may be because some variation in the faces that could also be synchronizing with the speech may be in the rigid head movements, which were removed from the facial data before performing the analyses. However, as noted before, the results do not appear to change significantly between the left singular analyses, which limit the dimensionality of both the speech and the faces, and the full analyses, which do not have this limit, suggesting that the results found may be robust to this differential, and the differential itself may not provide as much evidence towards or against synchrony as is being suggested here. In any case, the changes in curvatures between the CCA analyses still provide evidence towards turn-taking, and more research should be done to further explore these relationships.

After exploring the facial expressions and speech patterns of participants in dyadic conversations by using windowed cross-correlations and windowed canonical correlation

analyses, the methods used provided evidence towards the presence of turn-taking between facial expressions and paralinguistic speech patterns in conversations, as well as some potential evidence from the windowed cross-correlations towards the presence of synchrony. Both the evidence of turn-taking and synchrony have been found primarily in terms of speech in conversations, and are potentially driven by the designated speaker in the conversations. Although there is most likely synchrony happening with faces, or between faces and speech, any possible synchrony between faces was not found to be dependent on conversations, and was possibly masked by the overall effects of speech found in the analyses. In summary, although more research should be done on this topic, this project has shown promise in using both windowed cross-correlations and the novel method of windowed canonical correlation analyses to detect the presence of and further explore audiovisual relationships in dyadic conversations.

## 4.2 Limitations

Although these results seem promising at least for detecting turn-taking, it should be noted that there are some limitations with this study. First of all, the correlations calculated when performing WCC and CCA are linear, which might be ignoring any nonlinear relationships between the faces and the speech patterns. Also, because both the point coordinate and speech data is represented by so many variables (198 point coordinates and 136 frequencies per participant), the window size must be large enough so that there are more observations than variables. However, because of this window size, some effects that might be happening in a smaller time frame than the window size may be obscured. In addition, although for the surrogates the designated speakers were paired with designated listeners that did not include the original speaker-listener pairs, all participants in both the original and surrogate pairs came from the same experiment. While this allowed for more controlling of outside effects, this may have also attributed to similarities between the orig-

inal pairs and the surrogates as well. Next, while both participants were allowed to talk, the original experiment primed only one of the participants (the designated speaker) to do most of the talking. This helped with seeing more explicitly how designated speaking roles affect any potential turn-taking and synchrony, as well as how synchrony may occur in situations where one participant primarily talks. At the same time, it did not allow for as much insight into the presence of synchrony and turn-taking in everyday conversations, where the swapping of speaking roles is more frequent and natural. Finally, although differences between original pairs and the surrogates could be visually seen in the per-pair plots, these differences could not be detected when comparing distributions of WCC peak lags, CCA significant canonical pairs, and the corresponding CCA proportions of variance explained between the original pairs and the surrogates. This could be either because there are no effects of synchrony or because these distributions as a measure of synchrony are not sufficiently sensitive to detect these effects.

## 4.3 Future Directions and Impact

Overall, there is evidence towards the presence of turn-taking in the conversations, at least with regards to the amplitudes, or loudness, and frequencies, or pitch, of the conversations. However, most of this evidence was found visually when exploring individual results from the pairs. Therefore, I will explore more methods and techniques to measure synchrony in the future in order to detect these, and other audiovisual differences between the conversations.

Because the effect of speech, and more specifically the swapping of speaking roles, is so prominent in these conversations, any possible synchrony in the more nuanced aspects of the conversations, such as cadence, may not be noticeable. One way in which these nuances may be found is by choosing smaller windows. As noted before, a large window size was chosen in order to satisfy the restriction of needing more observations than vari-

ables. Although this larger window was required for the canonical correlation analyses that used the full matrices and was kept to maintain consistency across all the analyses, using decomposed data like the left singular matrices allow for smaller windows to be used in the analyses, which would allow faster changes in the conversations, such as role changes when both participants are talking, to be detected. This is because fewer variables would be needed to represent the data, and so fewer rows would be needed in the submatrices to satisfy this restriction.

Nuances in the conversation may also possibly be detected by changing how the data is decomposed. Therefore, more techniques to decompose the data will be considered in the future. One way that the speech data may be re-decomposed is by decomposing by using harmonics, rather than singular value decomposition. Because different harmonics have been known to include specific audio features, we can more explicitly explore these audio features by choosing which harmonics to include. Similarly, using audio feature extraction to decompose the speech data may shed light on more direct relationships between different audio features found in the speech. This would also allow for other types of paralinguistic measures to be included in the analyses.

In terms of facial movement, we can potentially test direct relationships between facial and audio features, such as between speech and mouth movement, by breaking down the facial expressions and movement into specific facial features. As mentioned previously, it may be further advantageous to compare rigid head movement to speech in addition to strictly nonrigid head movement, or facial expressions, as movement such as head nods may be more synchronous to speech inflections.

Another type of analysis that may help strengthen these claims would be to see how these results compare to the faces and speech patterns within a person, or within a speaking role. Previous research has shown that speech patterns and face and head movement tend to correlate well within a person, and so we can mediate the effect of speech and facial expressions between people by comparing the speech of a participant with the face of that

same participant in the conversation. Exploring these within-person relationships may also help with better understanding the concept of masking, or whether a person is communicating one emotion but feeling another. These periods of masking may also lead to changes in synchrony between people, and so masking would be important to test for and study in future research. We can also mediate the effect a particular designated speaking role has on the conversations by comparing designated speakers with other designated speakers, and designated listeners with designated listeners.

Other aspects of conversations should be considered in future studies, too, to see how the different parts of the conversation relate to each other. For example, since linguistics is a large component of speech, incorporating linguistic data into the analyses would also provide more insight towards the relationships between audiovisual dynamics in conversations. Several studies have shown evidence towards the relationship between affect and facial expressions and voice inflections, and so adding an affective or cognitive component would also provide a better understanding of the dynamics of the conversation. One way to do this would be to include physiological data, such as heartbeat data, that would indicate how much stress a person has. This would then help with understanding how stress in social situations, such as with affective disorders such as social anxiety, may play a role in conversations and communication as a whole.

Finally, by using data from other conversations, we can see how these results generalize to communication more broadly. For example, by doing this type of analyses with conversations between twins, we may see similar effects to what may be seen if these analyses were done in a within-person case, while providing enough variation in verbal and visual feedback that would be closer to a conversation between two people. This may help to provide a natural mediation that a within-person case would have, while keeping the between-person dynamics intact. Also, again, by performing these analyses on conversations that include someone with an affective disorder, or someone on the autism spectrum, we could again see how these conditions may affect conversation dynamics. Since only

one participant was primed to speak in the set of conversations that were used in the current study, performing these analyses on conversations with different speaking role dynamics, such as conversations where both participants are only whispering, conversations that have more than two people, and/or business meetings where there are clear distinctions between speakers and listeners, would provide more insight, as well, into differences between types of conversations. Overall, this research demonstrates potential ways that dynamics of communication can be broken down, and many avenues in which multivariate methods may help assist in better understanding the intricacies of these dynamics.

# References

Aamir, K. M., & Maud, M. A. (2007). Efficient spectral analysis of quasi stationary time series. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *1*(2).

Afifi, A., May, S., & Clark, V. A. (2003). *Computer-aided multivariate analysis*. CRC Press.

Arsalidou, M., Morris, D., & Taylor, M. J. (2011). Converging evidence for the advantage of dynamic facial expressions. *Brain Topography*, *24*(2), 149–163. Retrieved from http://dx.doi.org/10.1007/s10548-011-0171-4 doi: 10.1007/s10548-011-0171-4

Ashenfelter, K. T., Boker, S. M., Waddell, J. R., & Vitanov, N. (2009, Aug). Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. *J Exp Psychol Hum Percept Perform*, *35*(4), 1072-91. doi: 10.1037/a0015017

Avargel, Y., & Cohen, I. (2007, May). On multiplicative transfer function approximation in the short-time fourier transform domain. *IEEE Signal Processing Letters*, *14*(5), 337-340. doi: 10.1109/LSP.2006.888292

Boker, S. M., & Cohn, J. F. (2009). Dynamic faces: Insights from experiments and computation. In C. Curio, H. H. Bülthoff, & M. A. Giese (Eds.), (chap. Real-time dissociation

of facial appearance and dynamics during natural conversation). MIT Press Scholarship Online: August 2013.

Boker, S. M., Cohn, J. F., Theobald, B.-J., Matthews, I., Brick, T. R., & Spies, J. R. (2009, Dec). Effects of damping head movement and facial expression in dyadic conversation using real-time facial expression tracking and synthesized avatars. *Philos Trans R Soc Lond B Biol Sci*, *364*(1535), 3485-95. doi: 10.1098/rstb.2009.0152

Boker, S. M., Cohn, J. F., Theobald, B.-J., Matthews, I., Mangini, M., Spies, J. R., ... Brick, T. R. (2011, Jun). Something in the way we move: Motion dynamics, not perceived sex, influence head movements in conversation. *J Exp Psychol Hum Percept Perform*, *37*(3), 874-91. doi: 10.1037/a0021928

Boker, S. M., Xu, M., Rotondo, J. L., & King, K. (2002, Sep). Windowed cross–correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, *7*(3), 338-355.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., ... Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Icmi.*

Cattaneo, L., & Rizzolatti, G. (2009). The mirror neuron system. *Neurological Review*, *66*(5), 557-560.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *23*(6), 681–685.

Darrow, A.-A. (1990). The role of hearing in understanding music. *Music Educators Journal*, *77*(4), 24-27.

de Boer, E. (1980). Auditory physics. physical principles in hearing theory. i. *Physics Reports*, *62*(2), 87 - 174. Retrieved from `http://www.sciencedirect.com/`

science/article/pii/0370157380901003 doi: https://doi.org/10.1016/0370-1573(80)90100-3

de Boer, E. (1991). Auditory physics. physical principles in hearing theory. iii. *Physics Reports*, *203*(3), 125 - 231. Retrieved from http://www.sciencedirect.com/science/article/pii/037015739190068W doi: https://doi.org/10.1016/0370-1573(91)90068-W

Fan, N., Balan, R., & Rosca, J. (2004). Comparison of wavelet and fft based single channel speech signal noise reduction techniques. In *Wavelet applications in industrial processing ii.*

Fujiwara, K., & Daibo, I. (2016). Evaluating interpersonal synchrony: Wavelet transform toward an unstructured conversation. *Frontiers in Psychology*, *7*, 516. Retrieved from http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00516 doi: 10.3389/fpsyg.2016.00516

Ghazanfar, A. A., & Takahashi, D. Y. (2014, Jun). Facial expressions and the evolution of the speech rhythm. *Journal of Cognitive Neuroscience*, *26*(6), 1196–1207. doi: 10.1162/jocn_a_00575

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., & Rosenwald, D. P. (2014). Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, *32*(10), 641–647.

Heldner, M., & Edlund, J. (2010, Oct). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568. Retrieved from http://dx.doi.org/10.1016/j.wocn.2010.08.002 doi: 10.1016/j.wocn.2010.08.002

Honorof, D. N., & Whalen, D. H. (2010, Nov). Identification of speaker sex from one vowel across a range of fundamental frequencies. *The Journal of the Acoustical Society*

*of America*, *128*(5), 3095–3104. Retrieved from http://dx.doi.org/10.1121/1.3488347 doi: 10.1121/1.3488347

Hung, H., & Gatica-Perez, D. (2010, Oct). Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, *12*(6), 563-575. doi: 10.1109/TMM.2010.2055233

Iaci, R., Sriram, T., & Yin, X. (2010, Mar). Multivariate association and dimension reduction: A generalization of canonical correlation analysis. *Biometrics*, *66*(4), 1107–1118. Retrieved from http://dx.doi.org/10.1111/j.1541-0420.2010.01396.x doi: 10.1111/j.1541-0420.2010.01396.x

Iyengar, G., Nock, H. J., & Neti, C. (2003, July). Audio-visual synchrony for detection of monologues in video archives. In *Multimedia and expo, 2003. icme '03. proceedings. 2003 international conference on* (Vol. 1, p. I-329-32 vol.1). doi: 10.1109/ICME.2003.1220921

Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication*, *57*(Supplement C), 317 - 330. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167639313000691 doi: https://doi.org/10.1016/j.specom.2013.06.003

Krumhuber, E. G., Kappas, A., & Manstead, A. S. R. (2013, Jan). Effects of dynamic aspects of facial expressions: A review. *Emotion Review*, *5*(1), 41–46. Retrieved from http://dx.doi.org/10.1177/1754073912451349 doi: 10.1177/1754073912451349

Lieberman, P. (1975). Organization of behavior in face-to-face interaction. In A. Kendon, R. M. Harris, & M. R. Key (Eds.), (chap. Linguistic and Paralinguistic Interchange). Walter de Gruyter.

Melzer, T., Reiter, M., & Bischof, H. (2003, Sep). Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, *36*(9), 1961–1971. Retrieved from http://dx.doi.org/10.1016/S0031-3203(03)00058-X doi: 10.1016/s0031-3203(03)00058-x

Mowlaee, P., Saeidi, R., & Stylianou, Y. (2016). Advances in phase-aware signal processing in speech communication. *Speech Communication*, *81*, 1–29.

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *JOURNAL OF COMPUTING*, *2*(3).

Obermeier, C., & Gunter, T. C. (2015, Feb). Multisensory integration: The case of a time window of gesture–speech integration. *Journal of Cognitive Neuroscience*, *27*(2), 292–307. Retrieved from http://dx.doi.org/10.1162/jocn_a_00688 doi: 10.1162/jocn_a_00688

Paliwal, K. K., & Alsteris, L. (2003). Usefulness of phase spectrum in human speech perception. In *Eurospeech.*

Paliwal, K. K., Lyons, J. G., & Wojcicki, K. K. (2010). Preference for 20-40 ms window duration in speech analysis. *IEEE*.

Pearson, A. R., West, T. V., Dovidio, J. F., Powers, S. R., Buck, R., & Henning, R. (2008). The fragility of intergroup relations: Divergent effects of delayed audiovisual feedback in intergroup and intragroup interaction. *Psychological Science*, *19*(12), 1272–1279.

Rabiner, L. R., & Juang, B.-H. (1993). Fundamentals of speech recognition.

Ravindran, G., Shenbagadevi, S., & Selvam, V. S. (2010). Cepstral and linear prediction techniques for improving intelligibility and audibility of impaired speech. *Journal of*

*Biomedical Science and Engineering*, *03*(01), 85–94. Retrieved from http://dx.doi.org/10.4236/jbise.2010.31013 doi: 10.4236/jbise.2010.31013

Reyland, N. W. (2011). *Zbigniew preisner's three colors trilogy: Blue, white, red: A film score guide*. Scarecrow Press.

Sargin, M. E., Yemez, Y., Erzin, E., & Tekalp, A. M. (2007). Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, *9*(7), 1396–1403.

Sejdić, E., Jeffery, R., Kroonenberg, A. V., & Chau, T. (2012). An investigation of stride interval stationarity while listening to music or viewing television. *Human movement science*, *31*(3), 695–706.

Stanković, L., Orović, I., Stanković, S., & Amin, M. (2013). Compressive sensing based separation of nonstationary and stationary signals overlapping in time-frequency. *IEEE Transactions on Signal Processing*, *61*(18), 4562–4572.

Terven, J. R., Raducanu, B., de Luna, M. E. M., & Salas, J. (2016). Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices. *Neurocomputing*, *175*, 866-876.

Theobald, B.-J., Matthews, I., Mangini, M., Spies, J. R., Brick, T. R., Cohn, J. F., & Boker, S. M. (2009). Mapping and manipulating facial expression. *Language and Speech*, *52*, 369-386.

Tinati, M. A., & Mozaffary, B. (2005, October). Comparison of the wavelet and short time fourier transforms for spectral analysis of speech signals. In *5th wseas int. conf. on wavelet analysis and multirate systems* (p. 31-35).

Trentin, E., Scherer, S., & Schwenker, F. (2015). Emotion recognition from speech signals via a probabilistic echo-state network. *Pattern Recognition Letters*, *66*, 4–12.

Vatikiotis-Bateson, E., Barbosa, A. V., & Best, C. T. (2014). Articulatory coordination of two vocal tracts. *Journal of Phonetics*, *44*, 167–181.

Xiao, B., Georgiou, P. G., Lee, C.-C., Baucom, B., & Narayanan, S. S. (2013). Head motion synchrony and its correlation to affectivity in dyadic interactions. In *Multimedia and expo (icme)*.

Yegnanarayana, B., d'Alessandro, C., & Darsinos, V. (1998). An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Transactions on Speech and Audio processing*, *6*(1), 1–11.

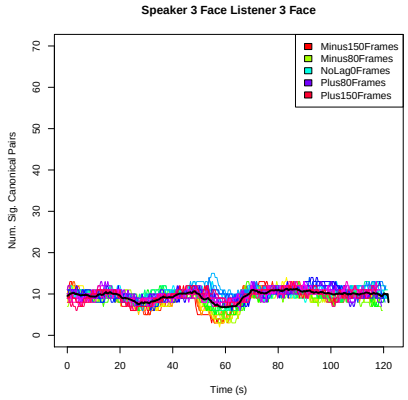
Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1), 23–43.

Zheng, W., Zhou, X., Zou, C., & Zhao, L. (2006). Facial expression recognition using kernel canonical correlation analysis (kcca). *IEEE transactions on neural networks*, *17*(1), 233–238.

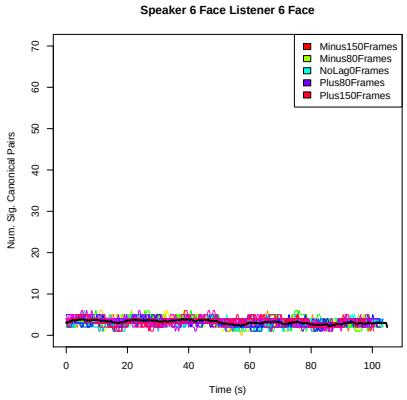
# A. WCCA Per-Pair Plots

## A.1 Full Analyses

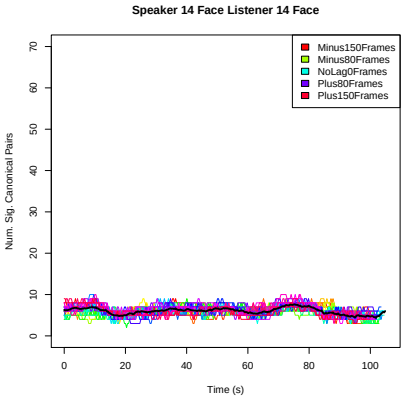
### A.1.1 Significant Canonical Pairs



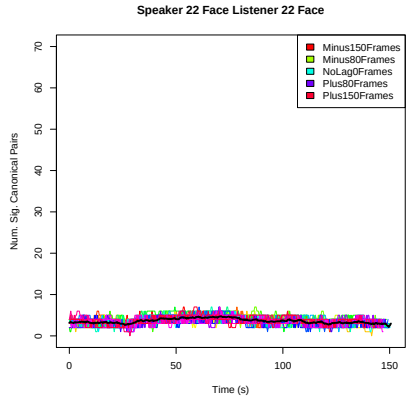

(a) Speaker Face Pair 3, Listener Face Pair 3




(b) Speaker Face Pair 6, Listener Face Pair 6

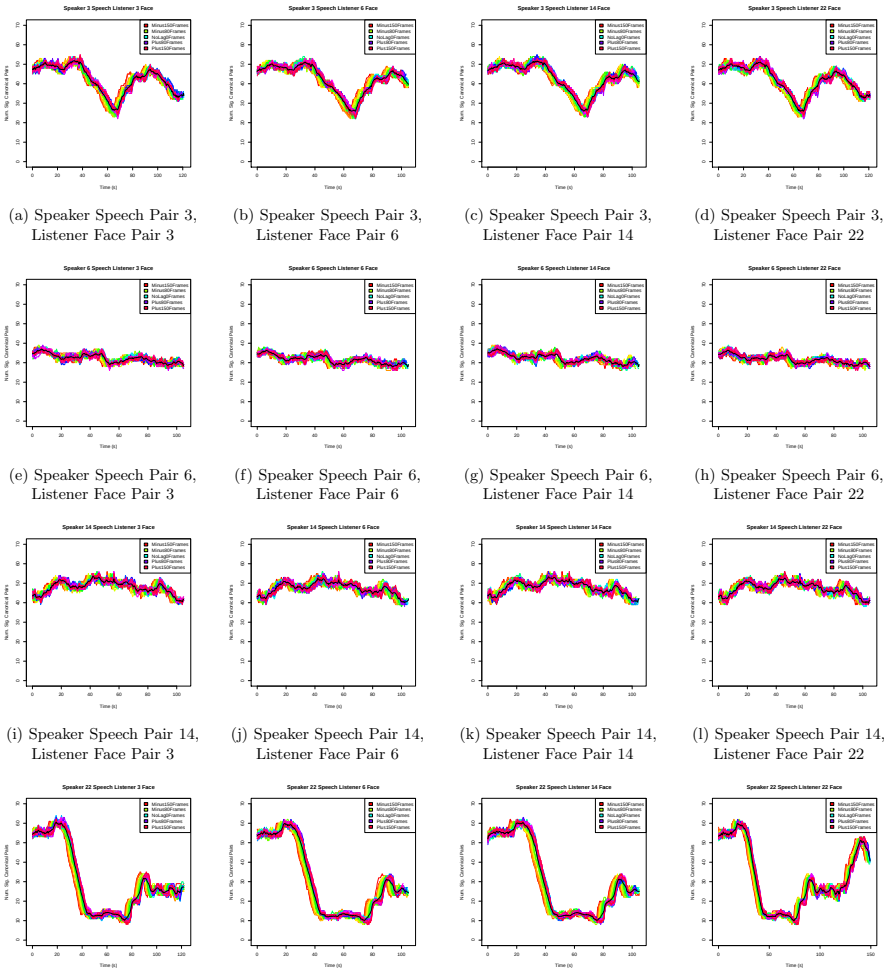



(c) Speaker Face Pair 14, Listener Face Pair 14




(d) Speaker Face Pair 22, Listener Face Pair 22

*Figure* A.1: Plots of the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's face for four representative pairs when using the full matrices. Note that surrogates for the Face-Face CCA analyses were only analyzed when using the left singular matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Speech Pair 3, Listener Face Pair 3

(b) Speaker Speech Pair 3, Listener Face Pair 6

(c) Speaker Speech Pair 3, Listener Face Pair 14

(d) Speaker Speech Pair 3, Listener Face Pair 22

(e) Speaker Speech Pair 6, Listener Face Pair 3

(f) Speaker Speech Pair 6, Listener Face Pair 6

(g) Speaker Speech Pair 6, Listener Face Pair 14

(h) Speaker Speech Pair 6, Listener Face Pair 22

(i) Speaker Speech Pair 14, Listener Face Pair 3

(j) Speaker Speech Pair 14, Listener Face Pair 6

(k) Speaker Speech Pair 14, Listener Face Pair 14

(l) Speaker Speech Pair 14, Listener Face Pair 22

(m) Speaker Speech Pair 22, Listener Face Pair 3

(n) Speaker Speech Pair 22, Listener Face Pair 6

(o) Speaker Speech Pair 22, Listener Face Pair 14

(p) Speaker Speech Pair 22, Listener Face Pair 22

*Figure* A.2: Plots of the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's face for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.

(a) Speaker Face Pair 3, Listener Speech Pair 3

(b) Speaker Face Pair 3, Listener Speech Pair 6

(c) Speaker Face Pair 3, Listener Speech Pair 14

(d) Speaker Face Pair 3, Listener Speech Pair 22

(e) Speaker Face Pair 6, Listener Speech Pair 3

(f) Speaker Face Pair 6, Listener Speech Pair 6

(g) Speaker Face Pair 6, Listener Speech Pair 14

(h) Speaker Face Pair 6, Listener Speech Pair 22

(i) Speaker Face Pair 14, Listener Speech Pair 3

(j) Speaker Face Pair 14, Listener Speech Pair 6

(k) Speaker Face Pair 14, Listener Speech Pair 14

(l) Speaker Face Pair 14, Listener Speech Pair 22

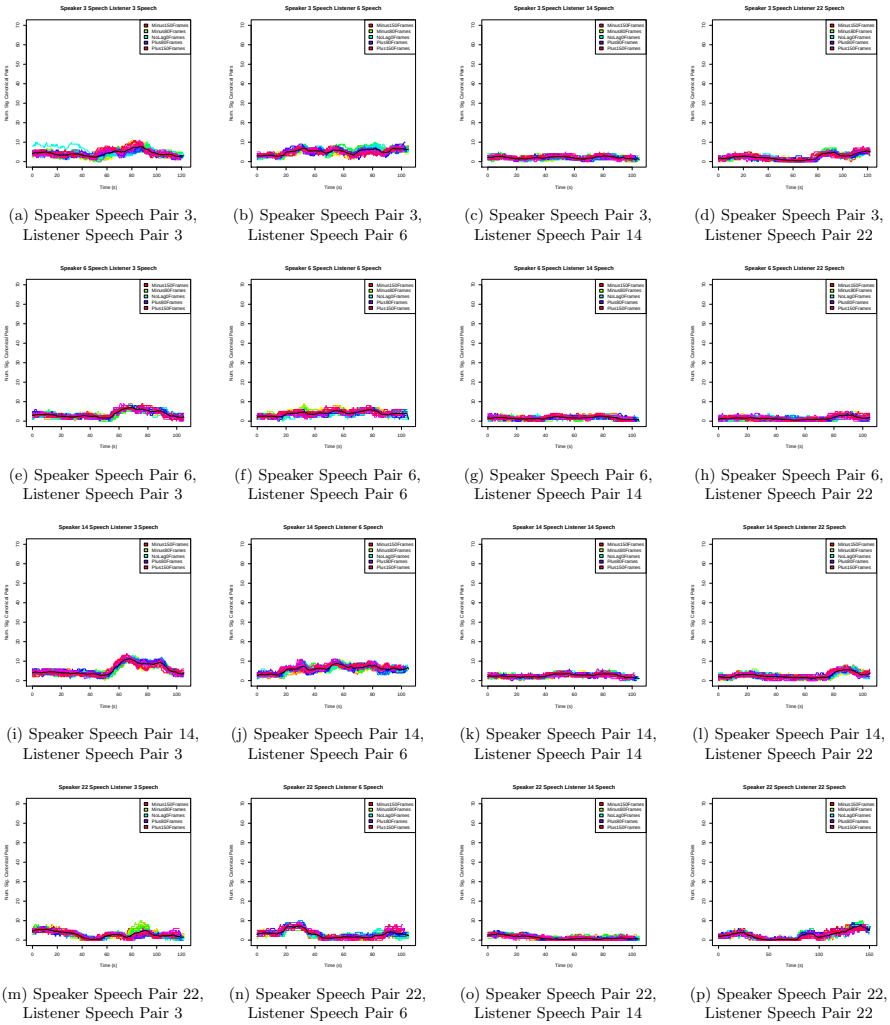
(m) Speaker Face Pair 22, Listener Speech Pair 3

(n) Speaker Face Pair 22, Listener Speech Pair 6

(o) Speaker Face Pair 22, Listener Speech Pair 14

(p) Speaker Face Pair 22, Listener Speech Pair 22

*Figure* A.3: Plots of the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's speech for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Speech Pair 3, Listener Speech Pair 3

(b) Speaker Speech Pair 3, Listener Speech Pair 6

(c) Speaker Speech Pair 3, Listener Speech Pair 14

(d) Speaker Speech Pair 3, Listener Speech Pair 22

(e) Speaker Speech Pair 6, Listener Speech Pair 3

(f) Speaker Speech Pair 6, Listener Speech Pair 6

(g) Speaker Speech Pair 6, Listener Speech Pair 14

(h) Speaker Speech Pair 6, Listener Speech Pair 22

(i) Speaker Speech Pair 14, Listener Speech Pair 3

(j) Speaker Speech Pair 14, Listener Speech Pair 6

(k) Speaker Speech Pair 14, Listener Speech Pair 14

(l) Speaker Speech Pair 14, Listener Speech Pair 22

(m) Speaker Speech Pair 22, Listener Speech Pair 3

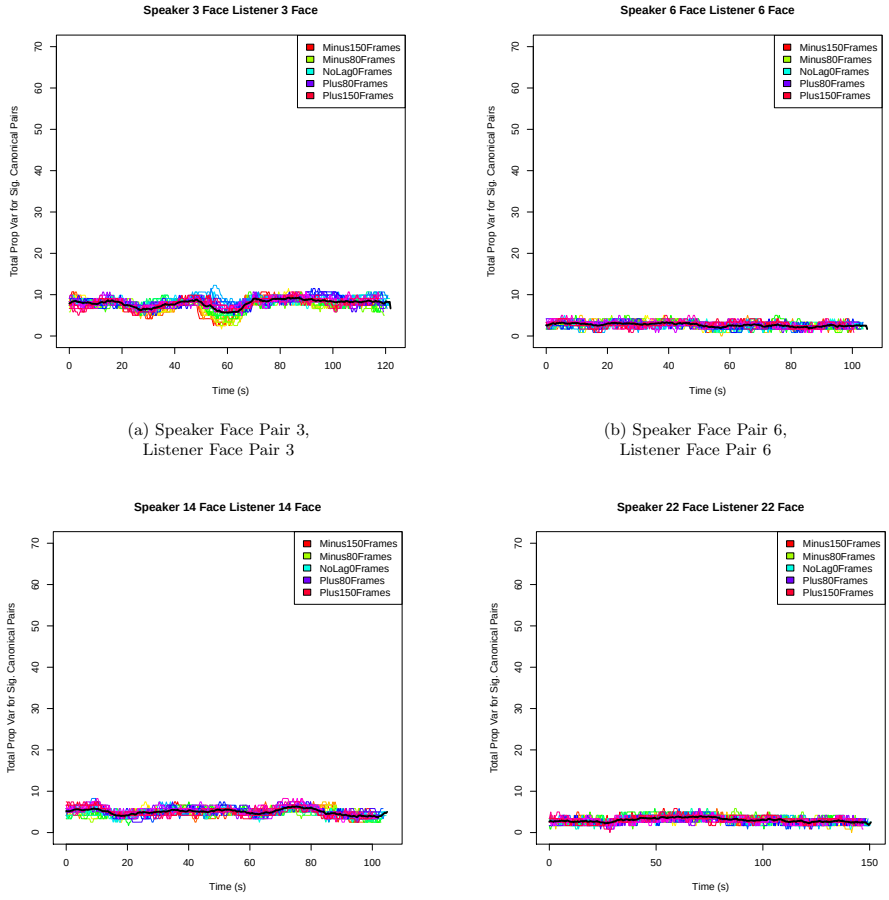
(n) Speaker Speech Pair 22, Listener Speech Pair 6

(o) Speaker Speech Pair 22, Listener Speech Pair 14

(p) Speaker Speech Pair 22, Listener Speech Pair 22

*Figure* A.4: Plots of the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's speech for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.

### A.1.2 Cumulative or Total Proportion Explained



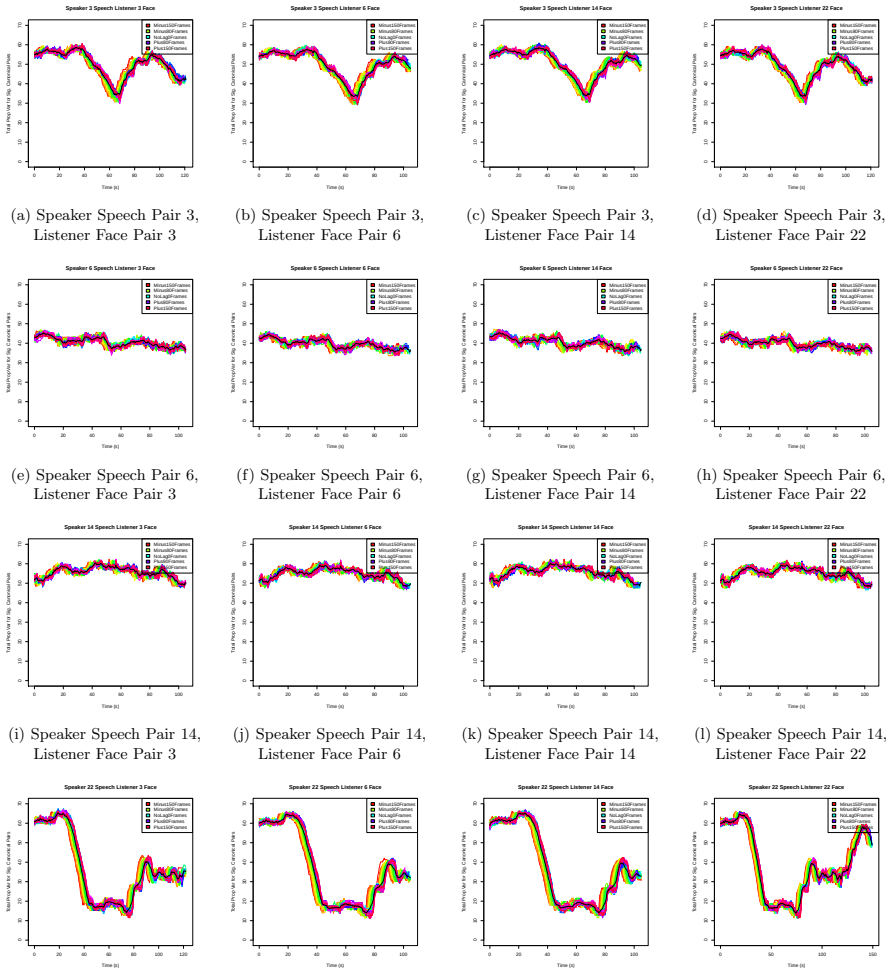

(a) Speaker Face Pair 3, Listener Face Pair 3

(b) Speaker Face Pair 6, Listener Face Pair 6

(c) Speaker Face Pair 14, Listener Face Pair 14

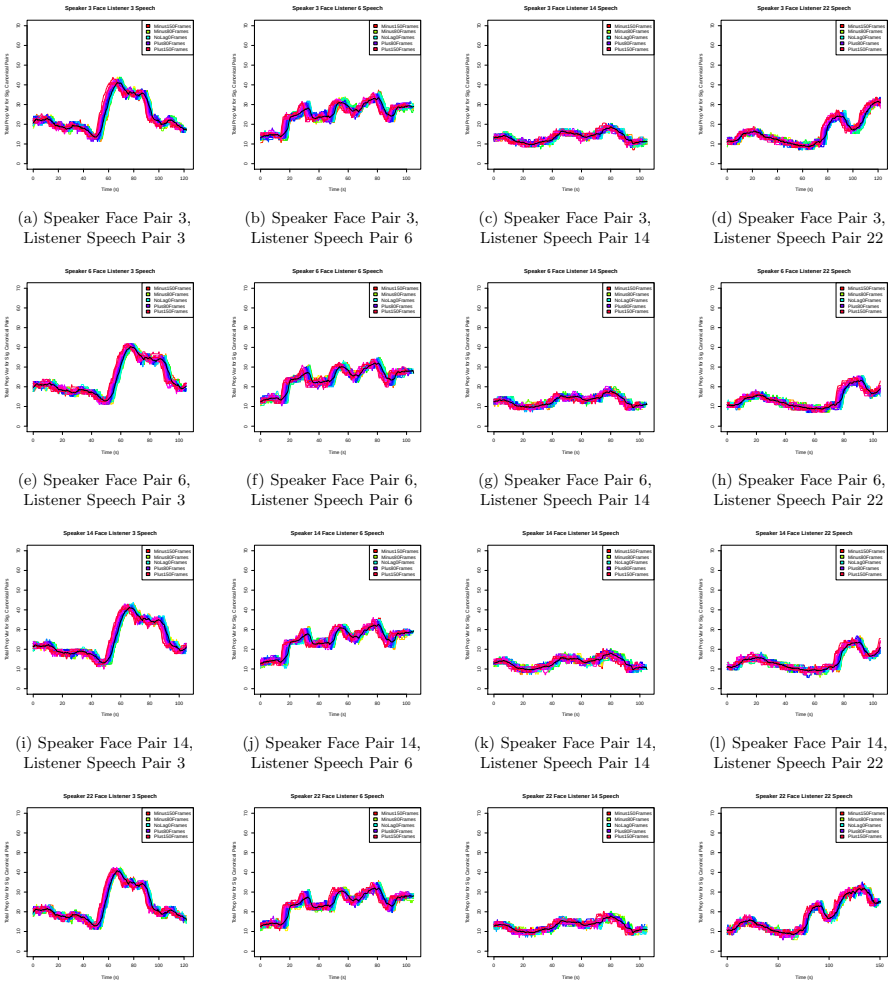
(d) Speaker Face Pair 22, Listener Face Pair 22

*Figure* A.5: Plots of the cumulative, or total, proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's face for four representative pairs when using the full matrices. Note that surrogates for the Face-Face CCA analyses were only analyzed when using the left singular matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Speech Pair 3, Listener Face Pair 3
(b) Speaker Speech Pair 3, Listener Face Pair 6
(c) Speaker Speech Pair 3, Listener Face Pair 14
(d) Speaker Speech Pair 3, Listener Face Pair 22
(e) Speaker Speech Pair 6, Listener Face Pair 3
(f) Speaker Speech Pair 6, Listener Face Pair 6
(g) Speaker Speech Pair 6, Listener Face Pair 14
(h) Speaker Speech Pair 6, Listener Face Pair 22
(i) Speaker Speech Pair 14, Listener Face Pair 3
(j) Speaker Speech Pair 14, Listener Face Pair 6
(k) Speaker Speech Pair 14, Listener Face Pair 14
(l) Speaker Speech Pair 14, Listener Face Pair 22
(m) Speaker Speech Pair 22, Listener Face Pair 3
(n) Speaker Speech Pair 22, Listener Face Pair 6
(o) Speaker Speech Pair 22, Listener Face Pair 14
(p) Speaker Speech Pair 22, Listener Face Pair 22

*Figure* A.6: Plots of the cumulative, or total, proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's face for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Face Pair 3, Listener Speech Pair 3

(b) Speaker Face Pair 3, Listener Speech Pair 6

(c) Speaker Face Pair 3, Listener Speech Pair 14

(d) Speaker Face Pair 3, Listener Speech Pair 22

(e) Speaker Face Pair 6, Listener Speech Pair 3

(f) Speaker Face Pair 6, Listener Speech Pair 6

(g) Speaker Face Pair 6, Listener Speech Pair 14

(h) Speaker Face Pair 6, Listener Speech Pair 22

(i) Speaker Face Pair 14, Listener Speech Pair 3

(j) Speaker Face Pair 14, Listener Speech Pair 6

(k) Speaker Face Pair 14, Listener Speech Pair 14

(l) Speaker Face Pair 14, Listener Speech Pair 22

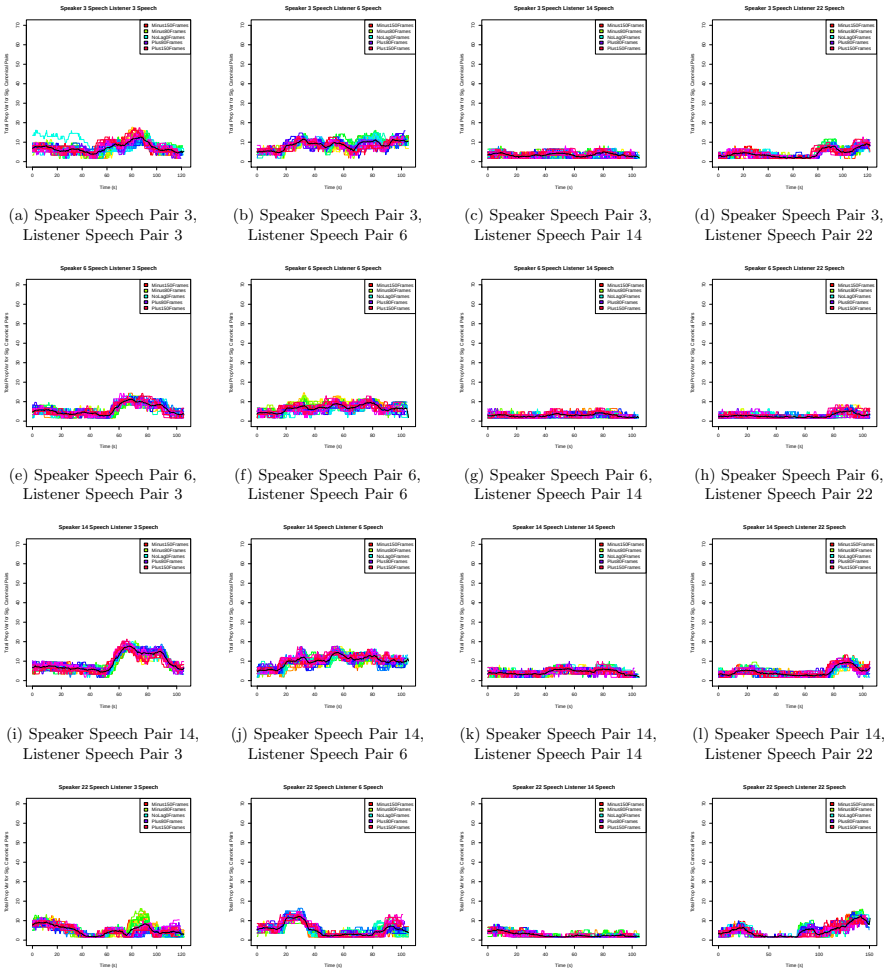
(m) Speaker Face Pair 22, Listener Speech Pair 3

(n) Speaker Face Pair 22, Listener Speech Pair 6

(o) Speaker Face Pair 22, Listener Speech Pair 14

(p) Speaker Face Pair 22, Listener Speech Pair 22

*Figure* A.7: Plots of the cumulative, or total, proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's speech for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Speech Pair 3, Listener Speech Pair 3

(b) Speaker Speech Pair 3, Listener Speech Pair 6

(c) Speaker Speech Pair 3, Listener Speech Pair 14

(d) Speaker Speech Pair 3, Listener Speech Pair 22

(e) Speaker Speech Pair 6, Listener Speech Pair 3

(f) Speaker Speech Pair 6, Listener Speech Pair 6

(g) Speaker Speech Pair 6, Listener Speech Pair 14

(h) Speaker Speech Pair 6, Listener Speech Pair 22

(i) Speaker Speech Pair 14, Listener Speech Pair 3

(j) Speaker Speech Pair 14, Listener Speech Pair 6

(k) Speaker Speech Pair 14, Listener Speech Pair 14

(l) Speaker Speech Pair 14, Listener Speech Pair 22

(m) Speaker Speech Pair 22, Listener Speech Pair 3

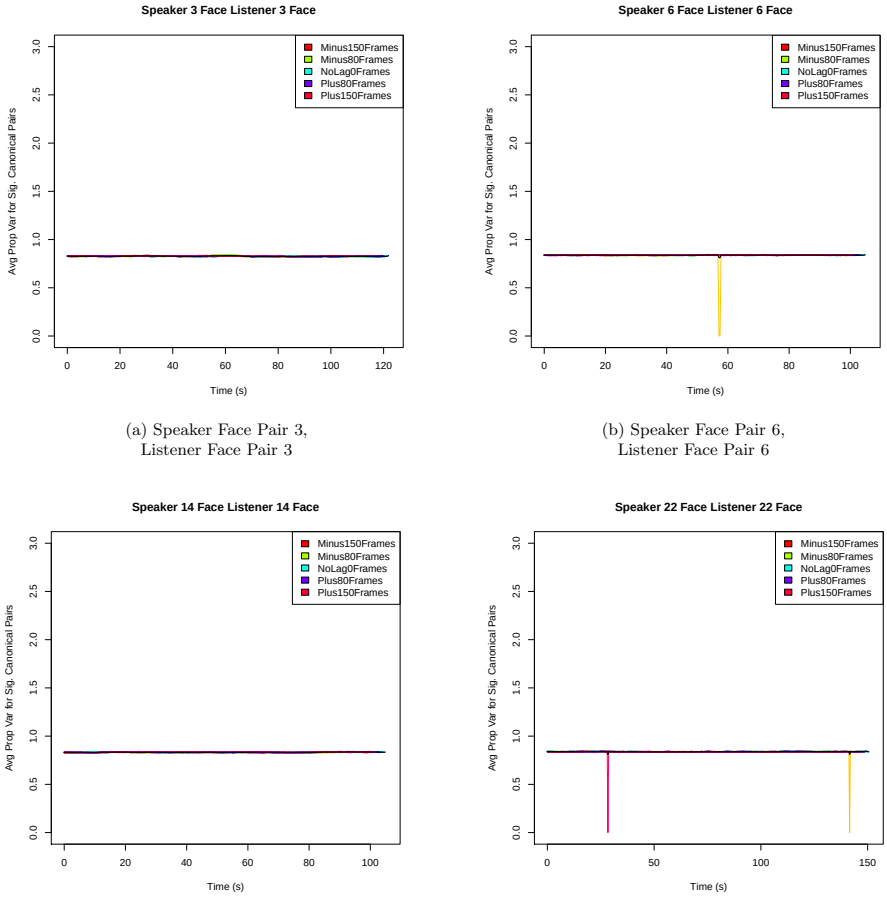
(n) Speaker Speech Pair 22, Listener Speech Pair 6

(o) Speaker Speech Pair 22, Listener Speech Pair 14

(p) Speaker Speech Pair 22, Listener Speech Pair 22

*Figure* A.8: Plots of the cumulative, or total, proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's speech for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.

### A.1.3 Average Proportion Explained



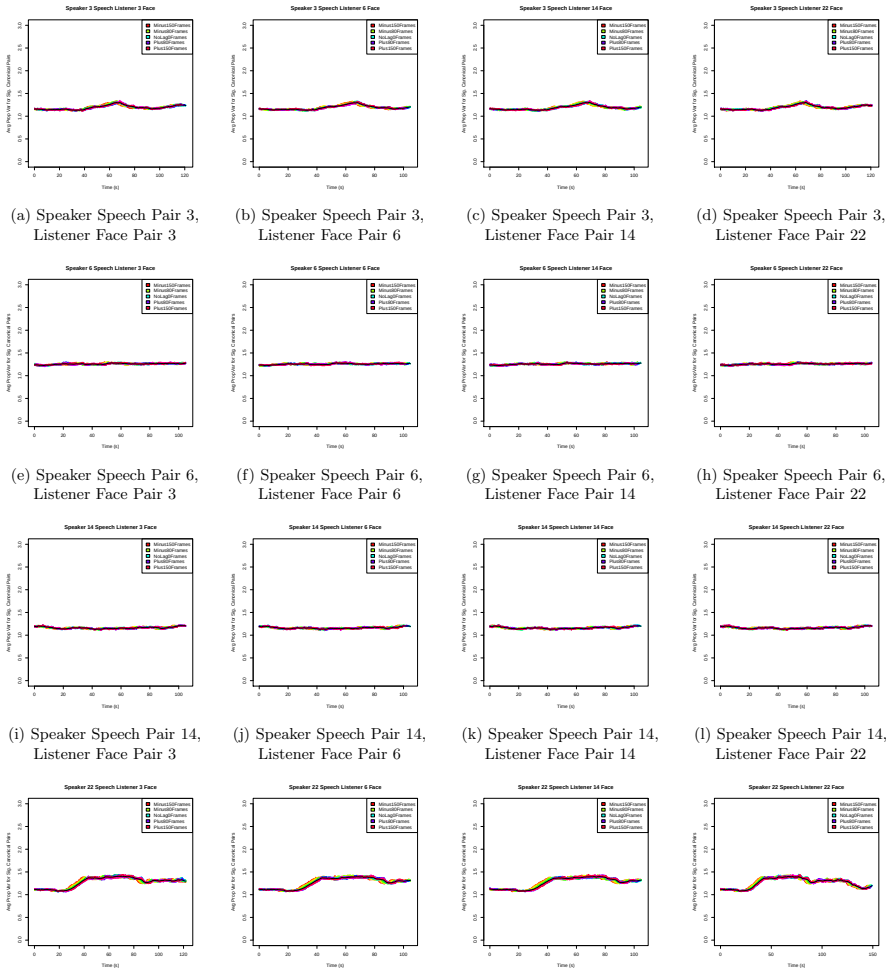

(a) Speaker Face Pair 3, Listener Face Pair 3

(b) Speaker Face Pair 6, Listener Face Pair 6

(c) Speaker Face Pair 14, Listener Face Pair 14

(d) Speaker Face Pair 22, Listener Face Pair 22

*Figure* A.9: Plots of the average proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's face for four representative pairs when using the full matrices. Note that surrogates for the Face-Face CCA analyses were only analyzed when using the left singular matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Speech Pair 3, Listener Face Pair 3

(b) Speaker Speech Pair 3, Listener Face Pair 6

(c) Speaker Speech Pair 3, Listener Face Pair 14

(d) Speaker Speech Pair 3, Listener Face Pair 22

(e) Speaker Speech Pair 6, Listener Face Pair 3

(f) Speaker Speech Pair 6, Listener Face Pair 6

(g) Speaker Speech Pair 6, Listener Face Pair 14

(h) Speaker Speech Pair 6, Listener Face Pair 22

(i) Speaker Speech Pair 14, Listener Face Pair 3

(j) Speaker Speech Pair 14, Listener Face Pair 6

(k) Speaker Speech Pair 14, Listener Face Pair 14

(l) Speaker Speech Pair 14, Listener Face Pair 22

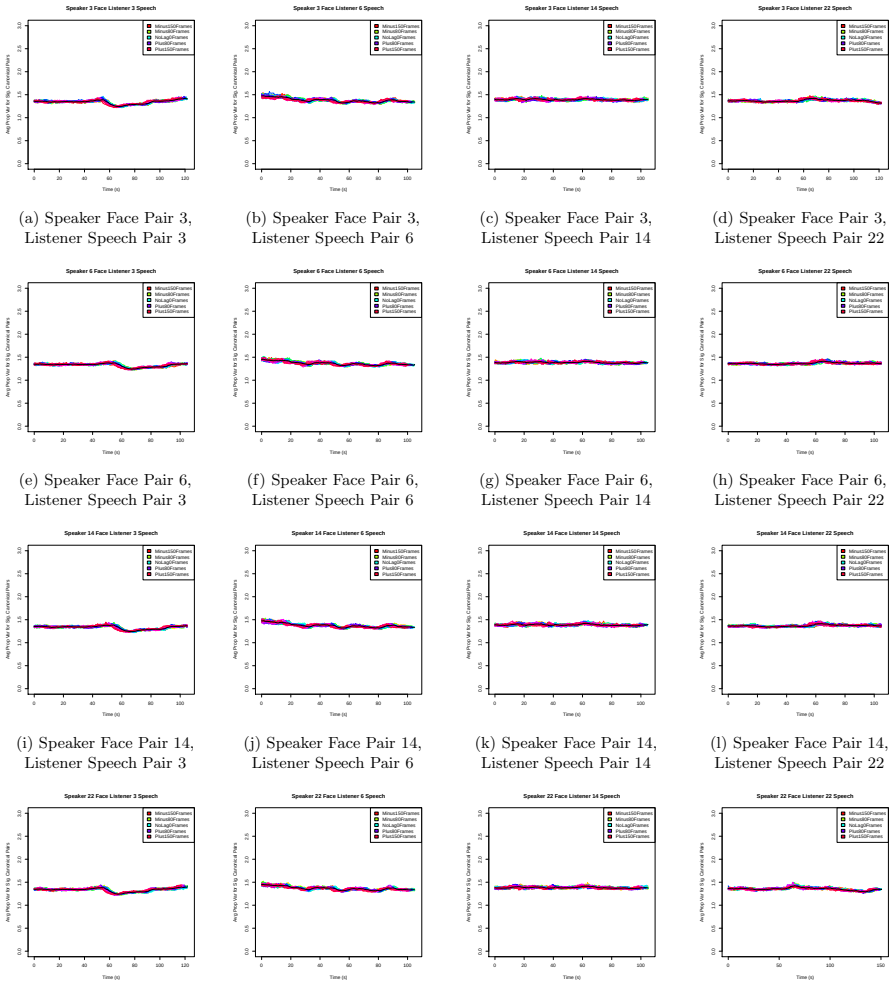
(m) Speaker Speech Pair 22, Listener Face Pair 3

(n) Speaker Speech Pair 22, Listener Face Pair 6

(o) Speaker Speech Pair 22, Listener Face Pair 14

(p) Speaker Speech Pair 22, Listener Face Pair 22

*Figure* A.10: Plots of the average proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's face for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Face Pair 3, Listener Speech Pair 3

(b) Speaker Face Pair 3, Listener Speech Pair 6

(c) Speaker Face Pair 3, Listener Speech Pair 14

(d) Speaker Face Pair 3, Listener Speech Pair 22

(e) Speaker Face Pair 6, Listener Speech Pair 3

(f) Speaker Face Pair 6, Listener Speech Pair 6

(g) Speaker Face Pair 6, Listener Speech Pair 14

(h) Speaker Face Pair 6, Listener Speech Pair 22

(i) Speaker Face Pair 14, Listener Speech Pair 3

(j) Speaker Face Pair 14, Listener Speech Pair 6

(k) Speaker Face Pair 14, Listener Speech Pair 14

(l) Speaker Face Pair 14, Listener Speech Pair 22

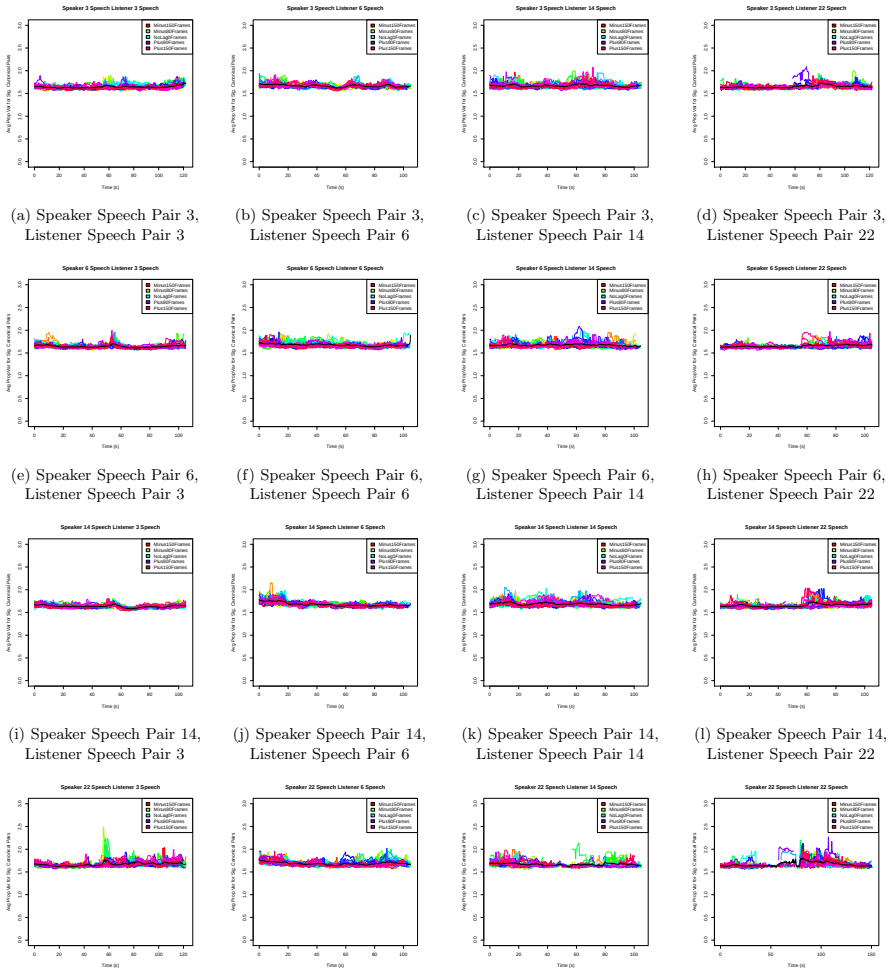
(m) Speaker Face Pair 22, Listener Speech Pair 3

(n) Speaker Face Pair 22, Listener Speech Pair 6

(o) Speaker Face Pair 22, Listener Speech Pair 14

(p) Speaker Face Pair 22, Listener Speech Pair 22

*Figure* A.11: Plots of the average proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's speech for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Speech Pair 3, Listener Speech Pair 3

(b) Speaker Speech Pair 3, Listener Speech Pair 6

(c) Speaker Speech Pair 3, Listener Speech Pair 14

(d) Speaker Speech Pair 3, Listener Speech Pair 22

(e) Speaker Speech Pair 6, Listener Speech Pair 3

(f) Speaker Speech Pair 6, Listener Speech Pair 6

(g) Speaker Speech Pair 6, Listener Speech Pair 14

(h) Speaker Speech Pair 6, Listener Speech Pair 22

(i) Speaker Speech Pair 14, Listener Speech Pair 3

(j) Speaker Speech Pair 14, Listener Speech Pair 6

(k) Speaker Speech Pair 14, Listener Speech Pair 14

(l) Speaker Speech Pair 14, Listener Speech Pair 22

(m) Speaker Speech Pair 22, Listener Speech Pair 3

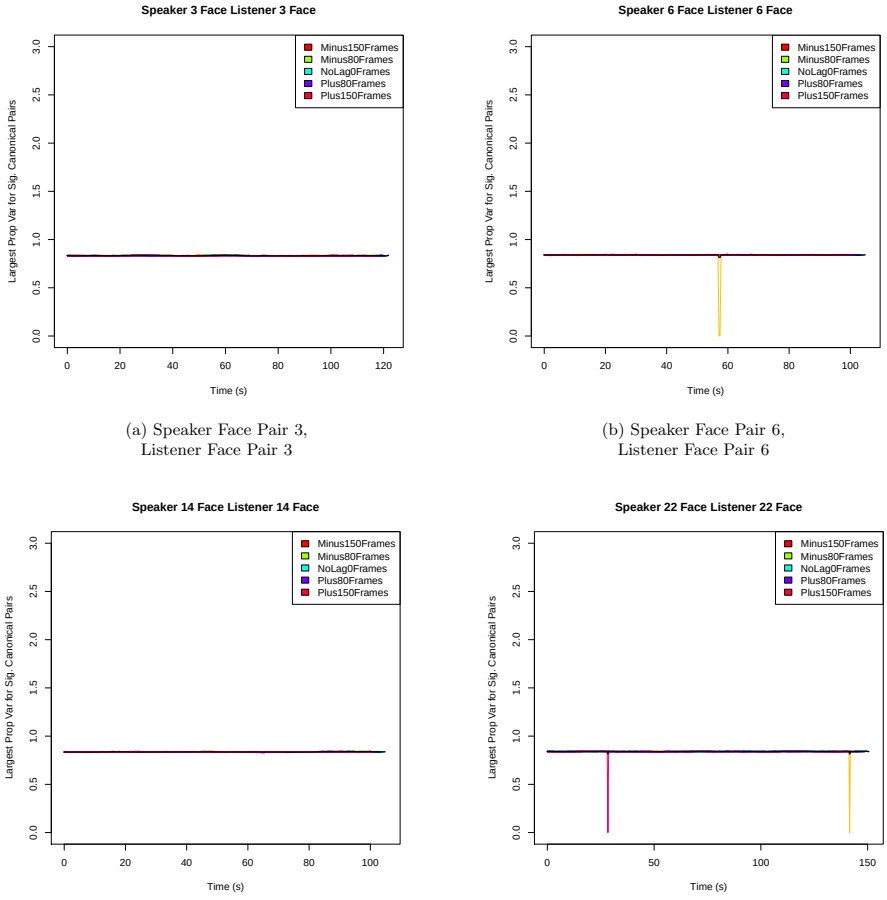
(n) Speaker Speech Pair 22, Listener Speech Pair 6

(o) Speaker Speech Pair 22, Listener Speech Pair 14

(p) Speaker Speech Pair 22, Listener Speech Pair 22

*Figure* A.12: Plots of the average proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's speech for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.

### A.1.4 Largest Proportion Explained



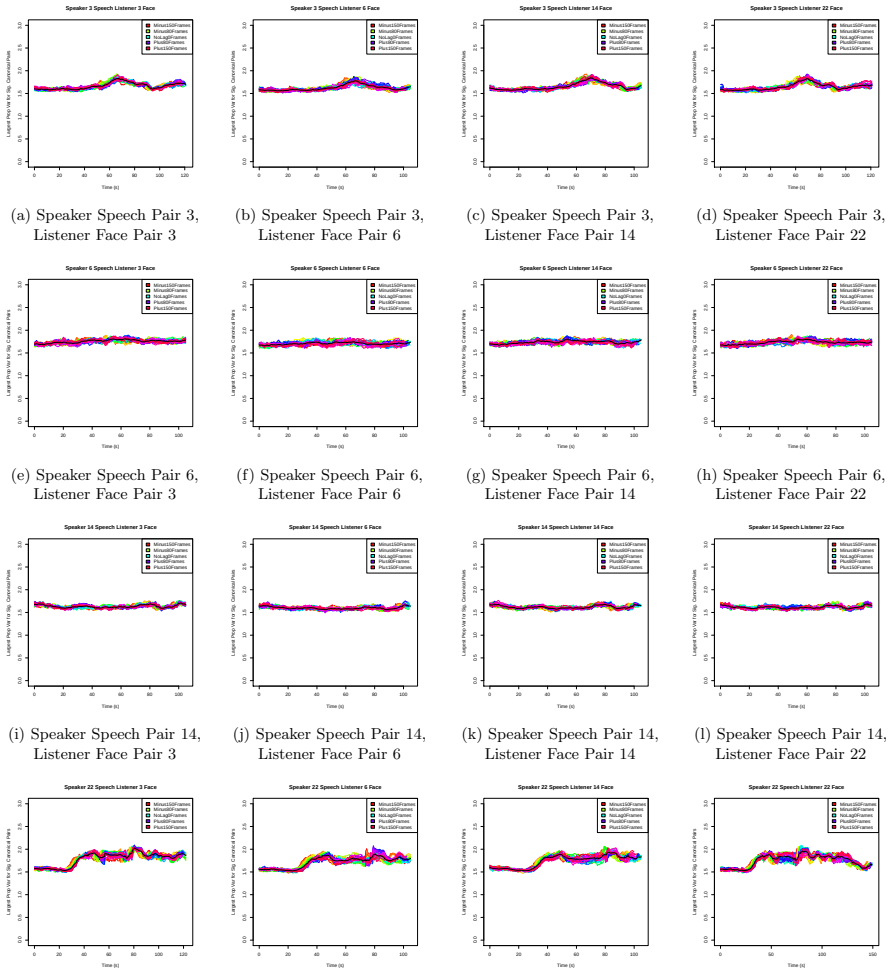

(a) Speaker Face Pair 3, Listener Face Pair 3

(b) Speaker Face Pair 6, Listener Face Pair 6

(c) Speaker Face Pair 14, Listener Face Pair 14

(d) Speaker Face Pair 22, Listener Face Pair 22

*Figure* A.13: Plots of the largest proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's face for four representative pairs when using the full matrices. Note that surrogates for the Face-Face CCA analyses were only analyzed when using the left singular matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Speech Pair 3, Listener Face Pair 3

(b) Speaker Speech Pair 3, Listener Face Pair 6

(c) Speaker Speech Pair 3, Listener Face Pair 14

(d) Speaker Speech Pair 3, Listener Face Pair 22

(e) Speaker Speech Pair 6, Listener Face Pair 3

(f) Speaker Speech Pair 6, Listener Face Pair 6

(g) Speaker Speech Pair 6, Listener Face Pair 14

(h) Speaker Speech Pair 6, Listener Face Pair 22

(i) Speaker Speech Pair 14, Listener Face Pair 3

(j) Speaker Speech Pair 14, Listener Face Pair 6

(k) Speaker Speech Pair 14, Listener Face Pair 14

(l) Speaker Speech Pair 14, Listener Face Pair 22

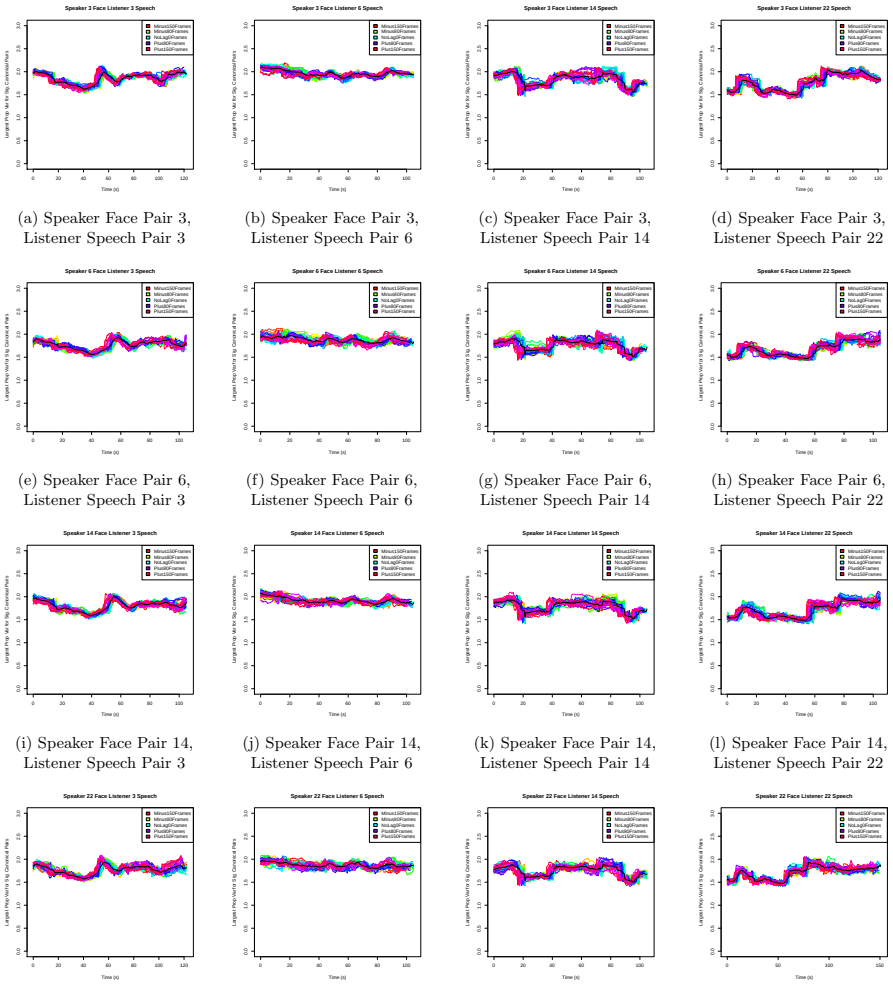
(m) Speaker Speech Pair 22, Listener Face Pair 3

(n) Speaker Speech Pair 22, Listener Face Pair 6

(o) Speaker Speech Pair 22, Listener Face Pair 14

(p) Speaker Speech Pair 22, Listener Face Pair 22

*Figure* A.14: Plots of the largest proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's face for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Face Pair 3, Listener Speech Pair 3

(b) Speaker Face Pair 3, Listener Speech Pair 6

(c) Speaker Face Pair 3, Listener Speech Pair 14

(d) Speaker Face Pair 3, Listener Speech Pair 22

(e) Speaker Face Pair 6, Listener Speech Pair 3

(f) Speaker Face Pair 6, Listener Speech Pair 6

(g) Speaker Face Pair 6, Listener Speech Pair 14

(h) Speaker Face Pair 6, Listener Speech Pair 22

(i) Speaker Face Pair 14, Listener Speech Pair 3

(j) Speaker Face Pair 14, Listener Speech Pair 6

(k) Speaker Face Pair 14, Listener Speech Pair 14

(l) Speaker Face Pair 14, Listener Speech Pair 22

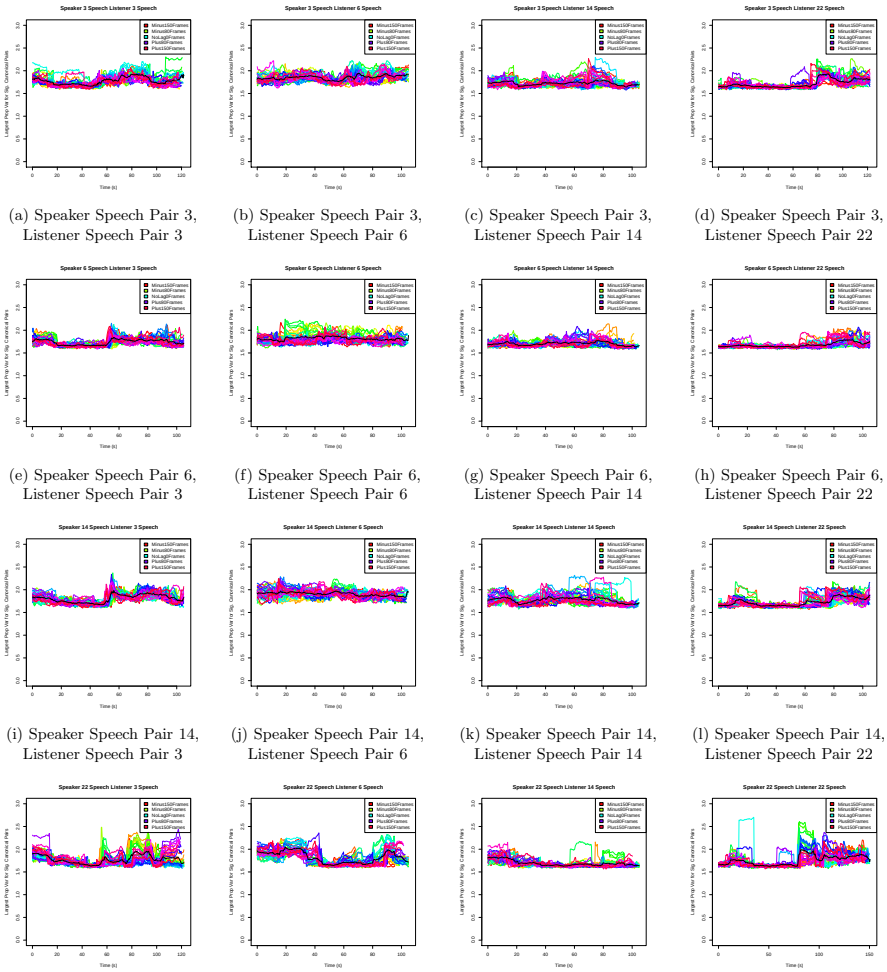
(m) Speaker Face Pair 22, Listener Speech Pair 3

(n) Speaker Face Pair 22, Listener Speech Pair 6

(o) Speaker Face Pair 22, Listener Speech Pair 14

(p) Speaker Face Pair 22, Listener Speech Pair 22

*Figure* A.15: Plots of the largest proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's speech for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



(a) Speaker Speech Pair 3, Listener Speech Pair 3

(b) Speaker Speech Pair 3, Listener Speech Pair 6

(c) Speaker Speech Pair 3, Listener Speech Pair 14

(d) Speaker Speech Pair 3, Listener Speech Pair 22

(e) Speaker Speech Pair 6, Listener Speech Pair 3

(f) Speaker Speech Pair 6, Listener Speech Pair 6

(g) Speaker Speech Pair 6, Listener Speech Pair 14

(h) Speaker Speech Pair 6, Listener Speech Pair 22

(i) Speaker Speech Pair 14, Listener Speech Pair 3

(j) Speaker Speech Pair 14, Listener Speech Pair 6

(k) Speaker Speech Pair 14, Listener Speech Pair 14

(l) Speaker Speech Pair 14, Listener Speech Pair 22

(m) Speaker Speech Pair 22, Listener Speech Pair 3

(n) Speaker Speech Pair 22, Listener Speech Pair 6

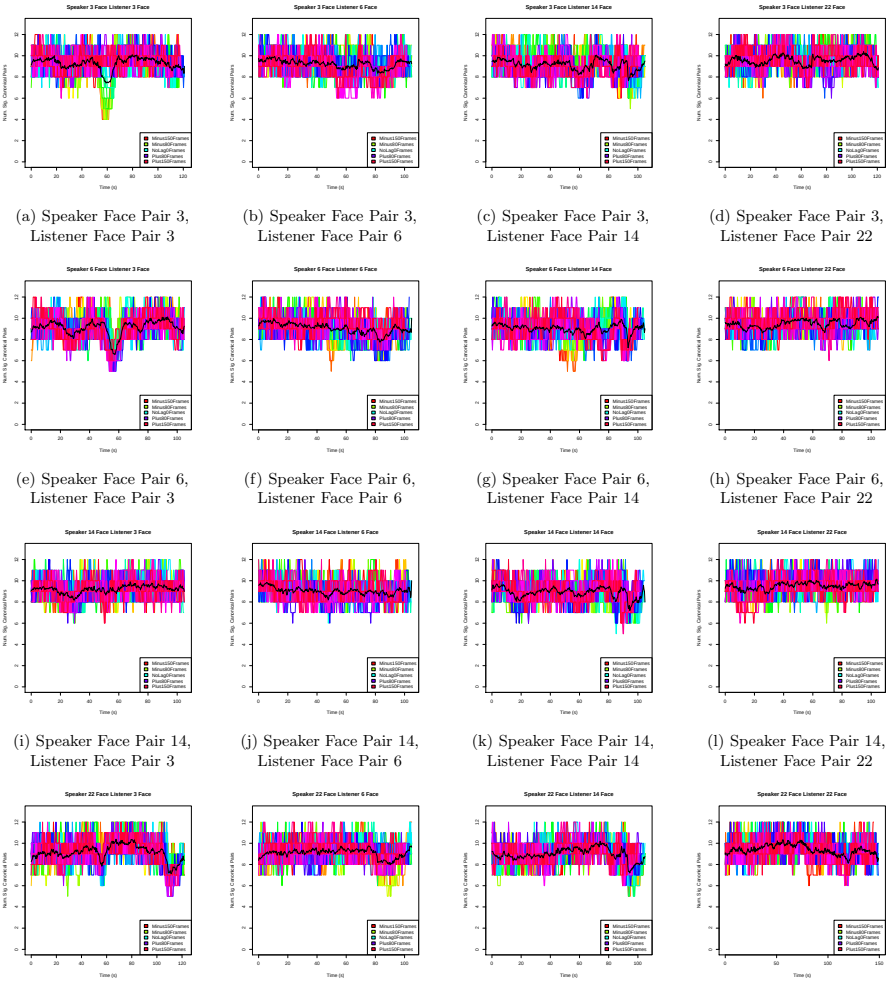
(o) Speaker Speech Pair 22, Listener Speech Pair 14

(p) Speaker Speech Pair 22, Listener Speech Pair 22

*Figure* A.16: Plots of the largest proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's speech for four representative pairs and their respective surrogates when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.

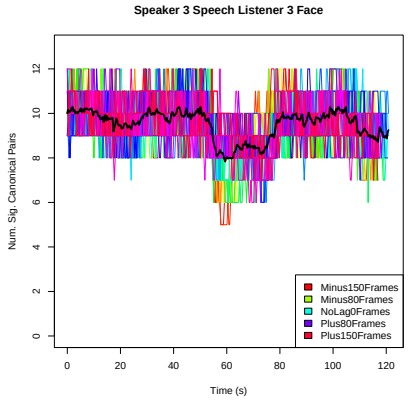
## A.2 Left Singular Analyses

### A.2.1 Significant Canonical Pairs

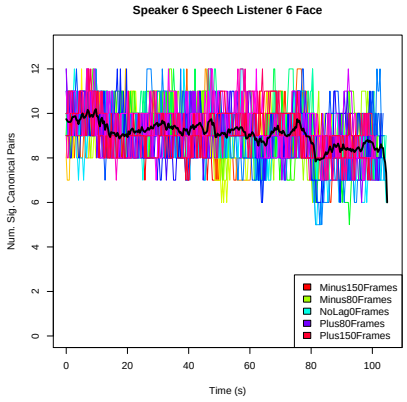


(a) Speaker Face Pair 3, Listener Face Pair 3
(b) Speaker Face Pair 3, Listener Face Pair 6
(c) Speaker Face Pair 3, Listener Face Pair 14
(d) Speaker Face Pair 3, Listener Face Pair 22
(e) Speaker Face Pair 6, Listener Face Pair 3
(f) Speaker Face Pair 6, Listener Face Pair 6
(g) Speaker Face Pair 6, Listener Face Pair 14
(h) Speaker Face Pair 6, Listener Face Pair 22
(i) Speaker Face Pair 14, Listener Face Pair 3
(j) Speaker Face Pair 14, Listener Face Pair 6
(k) Speaker Face Pair 14, Listener Face Pair 14
(l) Speaker Face Pair 14, Listener Face Pair 22
(m) Speaker Face Pair 22, Listener Face Pair 3
(n) Speaker Face Pair 22, Listener Face Pair 6
(o) Speaker Face Pair 22, Listener Face Pair 14
(p) Speaker Face Pair 22, Listener Face Pair 22

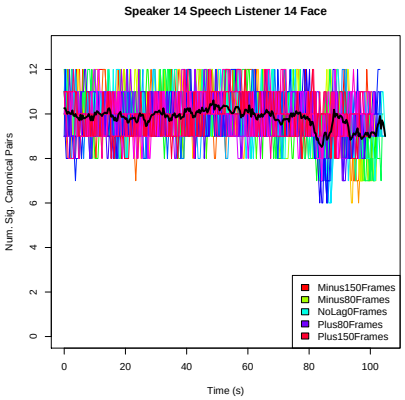
*Figure* A.17: Plots of the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's face for four representative pairs and their respective surrogates when using the left singular matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



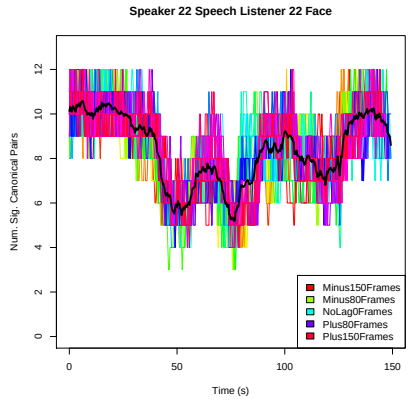

(a) Speaker Speech Pair 3, Listener Face Pair 3




(b) Speaker Speech Pair 6, Listener Face Pair 6

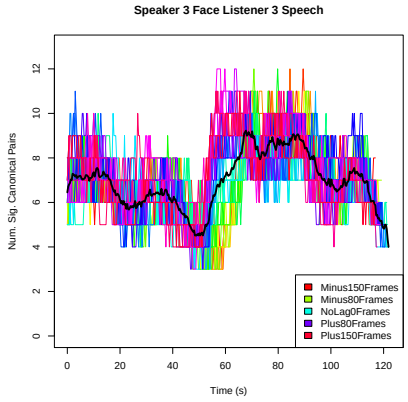



(c) Speaker Speech Pair 14, Listener Face Pair 14

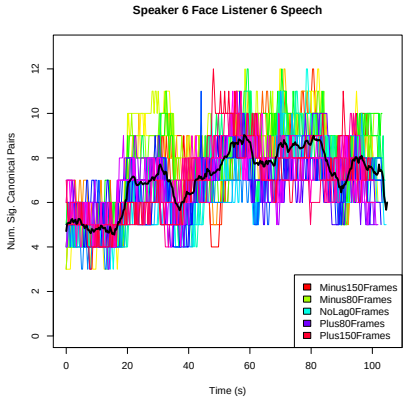



(d) Speaker Speech Pair 22, Listener Face Pair 22

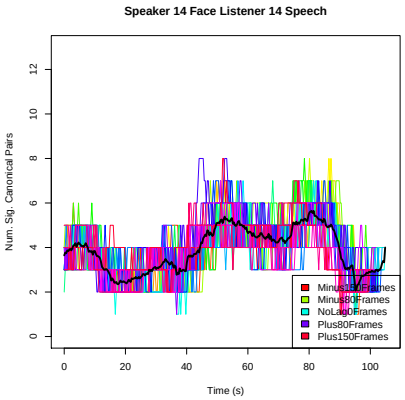
*Figure* A.18: Plots of the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's face for four representative pairs when using the left singular matrices. Note that surrogates for the Speech-Face CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



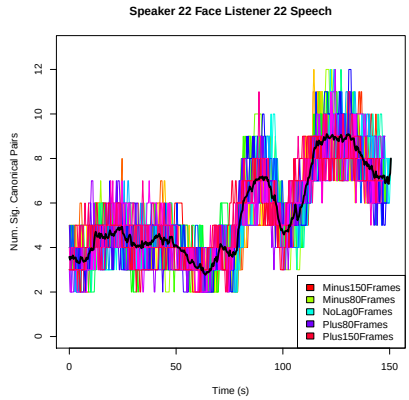

(a) Speaker Face Pair 3,
Listener Speech Pair 3




(b) Speaker Face Pair 6,
Listener Speech Pair 6

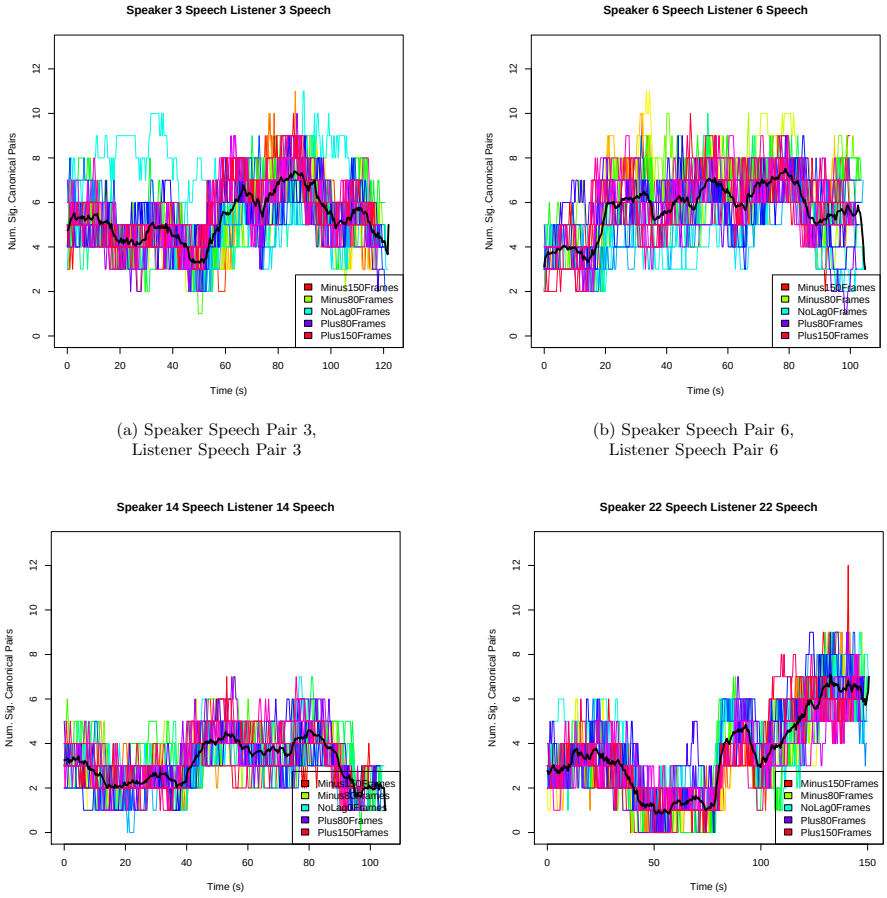



(c) Speaker Face Pair 14,
Listener Speech Pair 14




(d) Speaker Face Pair 22,
Listener Speech Pair 22

*Figure* A.19: Plots of the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's speech for four representative pairs when using the left singular matrices. Note that surrogates for the Face-Speech CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.




(a) Speaker Speech Pair 3, Listener Speech Pair 3

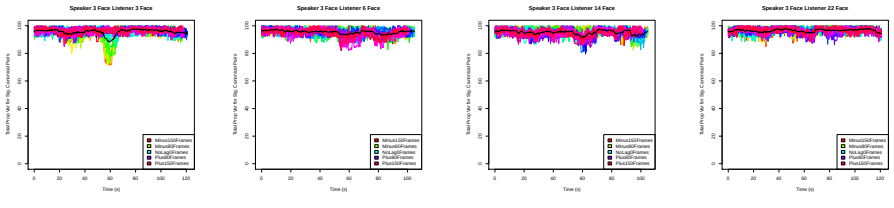
(b) Speaker Speech Pair 6, Listener Speech Pair 6

(c) Speaker Speech Pair 14, Listener Speech Pair 14

(d) Speaker Speech Pair 22, Listener Speech Pair 22

*Figure* A.20: Plots of the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's speech for four representative pairs when using the left singular matrices. Note that surrogates for the Speech-Speech CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.

### A.2.2 Cumulative or Total Proportion Explained



(a) Speaker Face Pair 3, Listener Face Pair 3



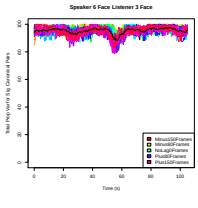
(b) Speaker Face Pair 3, Listener Face Pair 6



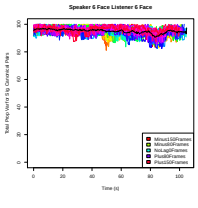
(c) Speaker Face Pair 3, Listener Face Pair 14



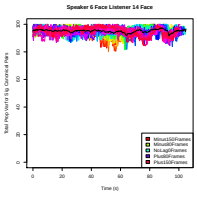
(d) Speaker Face Pair 3, Listener Face Pair 22



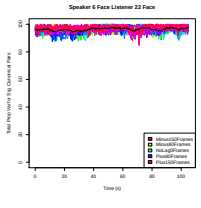
(e) Speaker Face Pair 6, Listener Face Pair 3



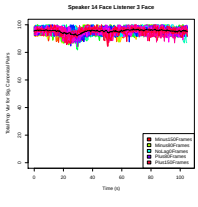
(f) Speaker Face Pair 6, Listener Face Pair 6



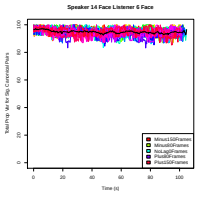
(g) Speaker Face Pair 6, Listener Face Pair 14



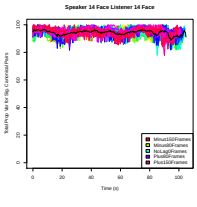
(h) Speaker Face Pair 6, Listener Face Pair 22



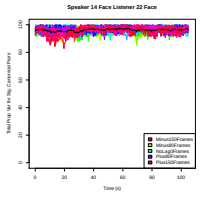
(i) Speaker Face Pair 14, Listener Face Pair 3



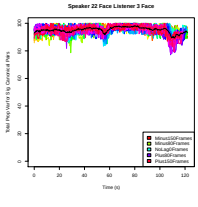
(j) Speaker Face Pair 14, Listener Face Pair 6



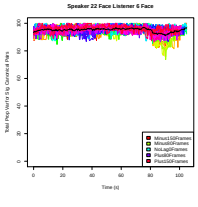
(k) Speaker Face Pair 14, Listener Face Pair 14



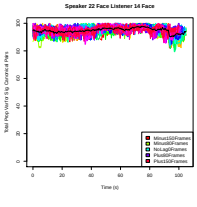
(l) Speaker Face Pair 14, Listener Face Pair 22



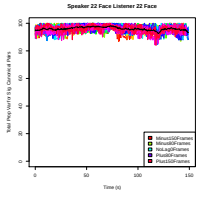
(m) Speaker Face Pair 22, Listener Face Pair 3



(n) Speaker Face Pair 22, Listener Face Pair 6

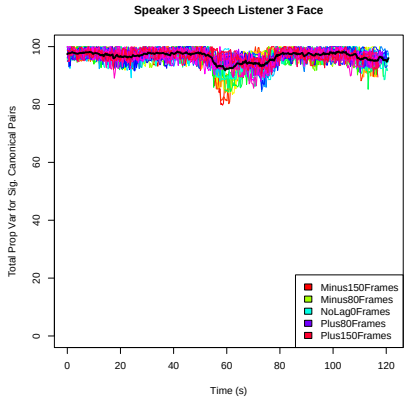


(o) Speaker Face Pair 22, Listener Face Pair 14

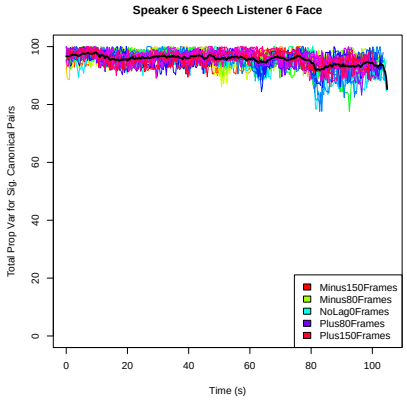


(p) Speaker Face Pair 22, Listener Face Pair 22

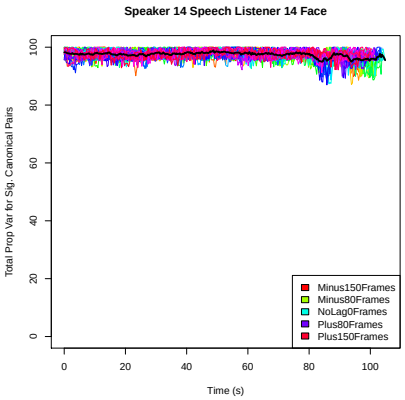
*Figure* A.21: Plots of the cumulative, or total, proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's face for four representative pairs and their respective surrogates when using the left singular matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



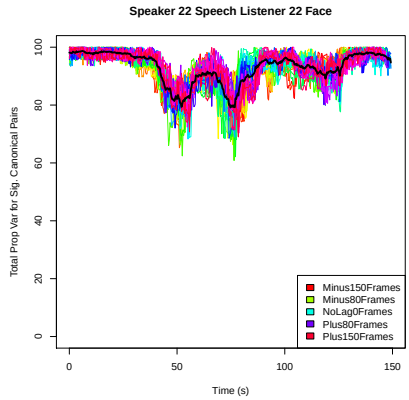

(a) Speaker Speech Pair 3, Listener Face Pair 3




(b) Speaker Speech Pair 6, Listener Face Pair 6

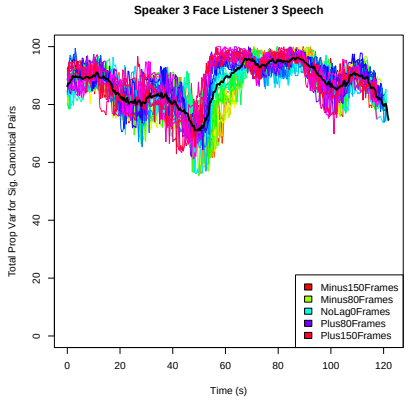



(c) Speaker Speech Pair 14, Listener Face Pair 14

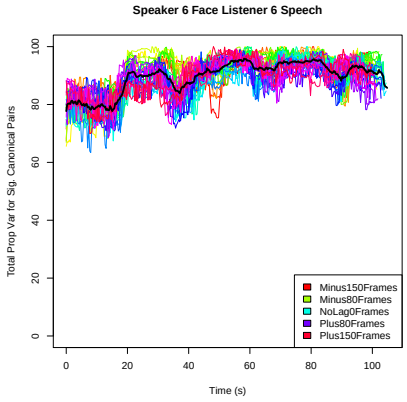



(d) Speaker Speech Pair 22, Listener Face Pair 22

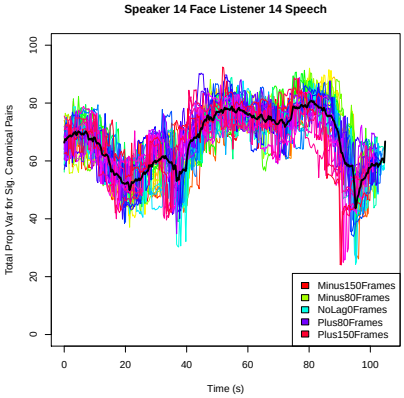
*Figure* A.22: Plots of the cumulative, or total, proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's face for four representative pairs when using the left singular matrices. Note that surrogates for the Speech-Face CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



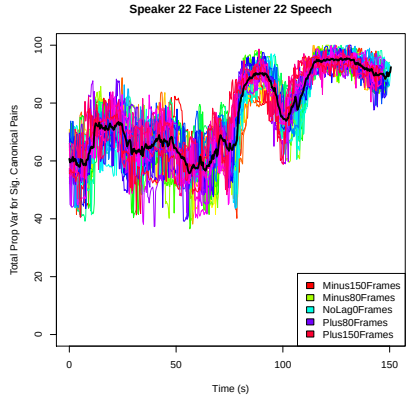

(a) Speaker Face Pair 3, Listener Speech Pair 3




(b) Speaker Face Pair 6, Listener Speech Pair 6

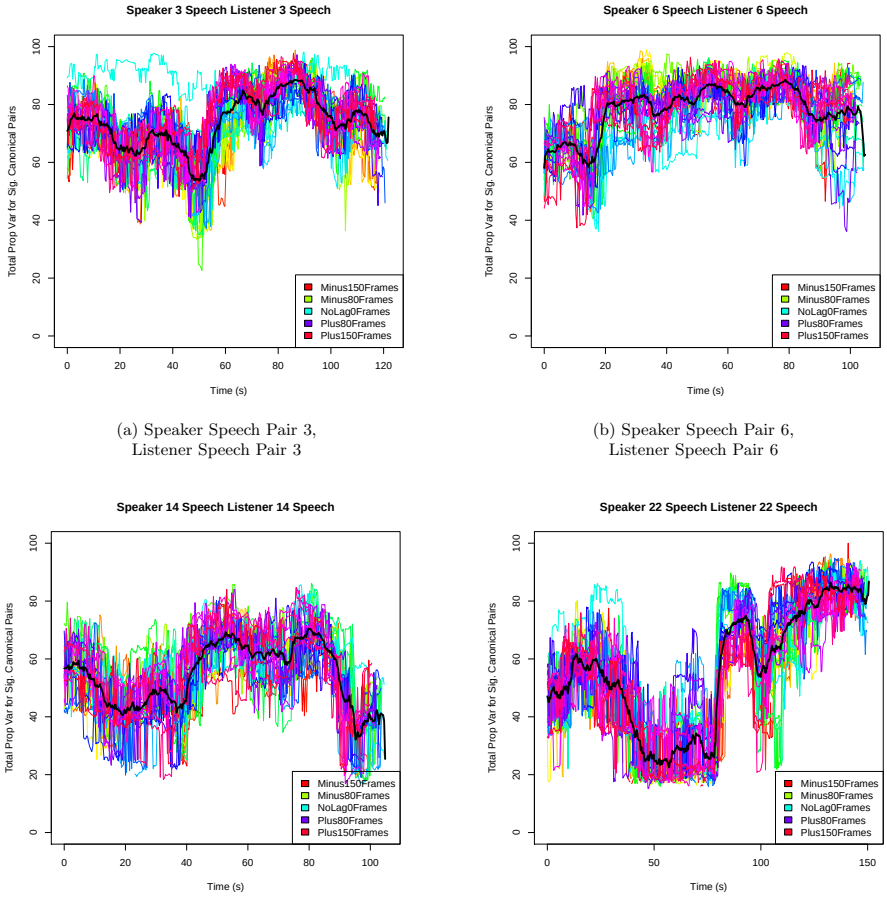



(c) Speaker Face Pair 14, Listener Speech Pair 14




(d) Speaker Face Pair 22, Listener Speech Pair 22

*Figure* A.23: Plots of the cumulative, or total, proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's speech for four representative pairs when using the left singular matrices. Note that surrogates for the Face-Speech CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.




(a) Speaker Speech Pair 3, Listener Speech Pair 3

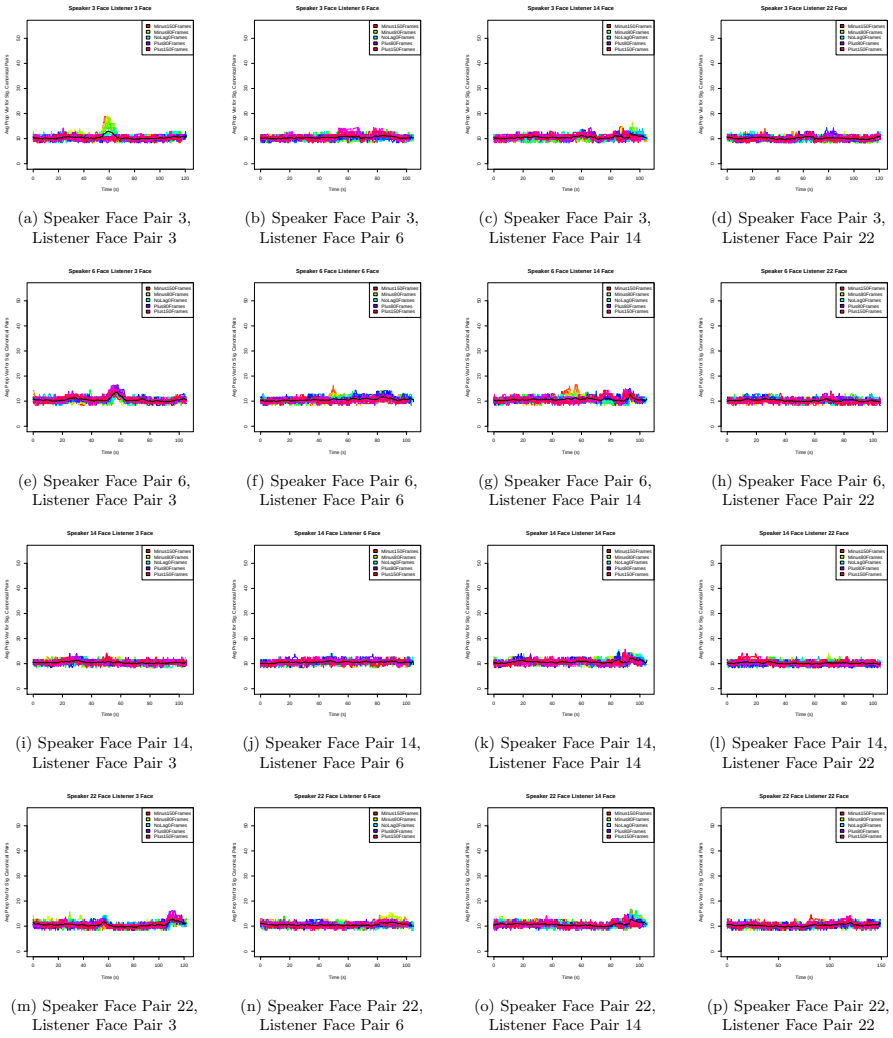
(b) Speaker Speech Pair 6, Listener Speech Pair 6

(c) Speaker Speech Pair 14, Listener Speech Pair 14

(d) Speaker Speech Pair 22, Listener Speech Pair 22

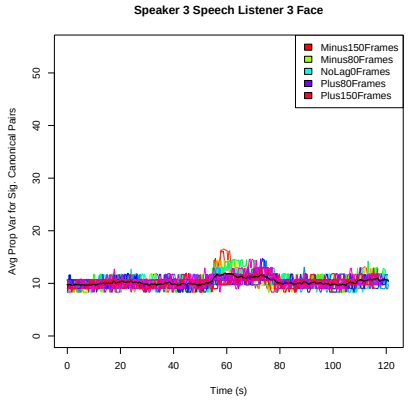
*Figure* A.24: Plots of the cumulative, or total, proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's speech for four representative pairs when using the left singular matrices. Note that surrogates for the Speech-Speech CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.

### A.2.3 Average Proportion Explained

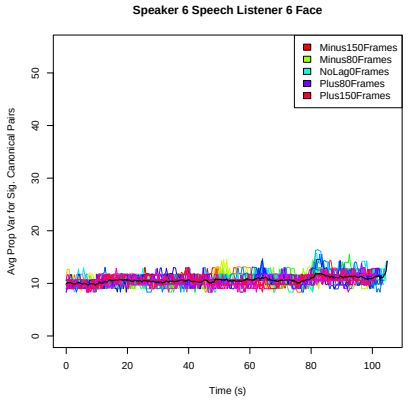


(a) Speaker Face Pair 3, Listener Face Pair 3
(b) Speaker Face Pair 3, Listener Face Pair 6
(c) Speaker Face Pair 3, Listener Face Pair 14
(d) Speaker Face Pair 3, Listener Face Pair 22
(e) Speaker Face Pair 6, Listener Face Pair 3
(f) Speaker Face Pair 6, Listener Face Pair 6
(g) Speaker Face Pair 6, Listener Face Pair 14
(h) Speaker Face Pair 6, Listener Face Pair 22
(i) Speaker Face Pair 14, Listener Face Pair 3
(j) Speaker Face Pair 14, Listener Face Pair 6
(k) Speaker Face Pair 14, Listener Face Pair 14
(l) Speaker Face Pair 14, Listener Face Pair 22
(m) Speaker Face Pair 22, Listener Face Pair 3
(n) Speaker Face Pair 22, Listener Face Pair 6
(o) Speaker Face Pair 22, Listener Face Pair 14
(p) Speaker Face Pair 22, Listener Face Pair 22

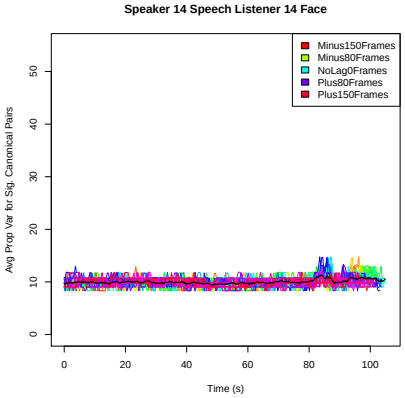
*Figure* A.25: Plots of the average proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's face for four representative pairs and their respective surrogates when using the left singular matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



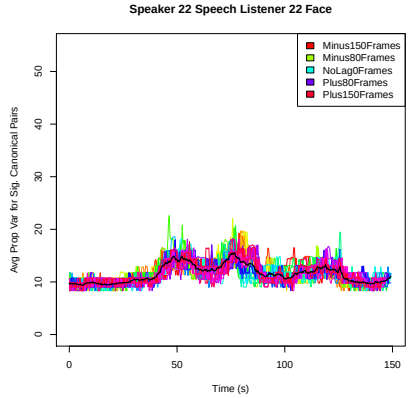

(a) Speaker Speech Pair 3, Listener Face Pair 3




(b) Speaker Speech Pair 6, Listener Face Pair 6

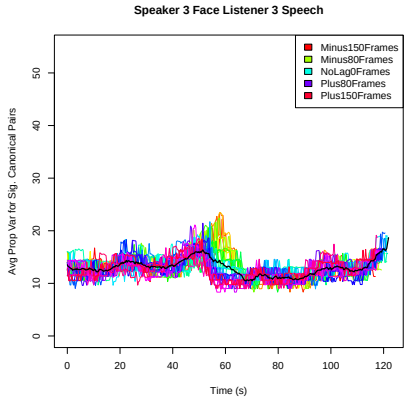



(c) Speaker Speech Pair 14, Listener Face Pair 14

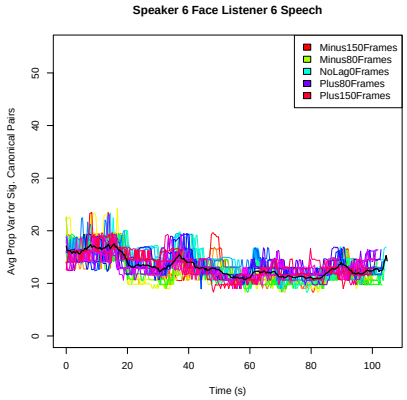



(d) Speaker Speech Pair 22, Listener Face Pair 22

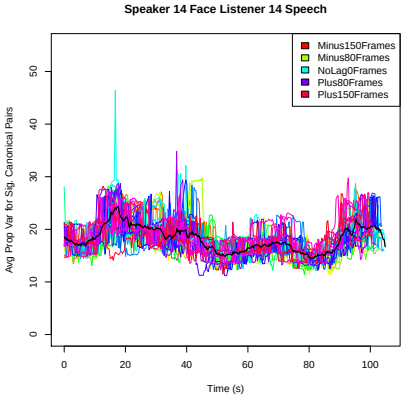
*Figure* A.26: Plots of the average proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's face for four representative pairs when using the left singular matrices. Note that surrogates for the Speech-Face CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



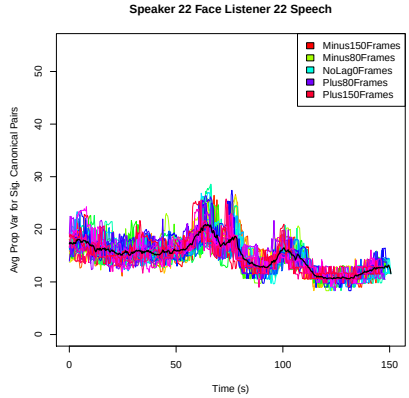

(a) Speaker Face Pair 3, Listener Speech Pair 3




(b) Speaker Face Pair 6, Listener Speech Pair 6

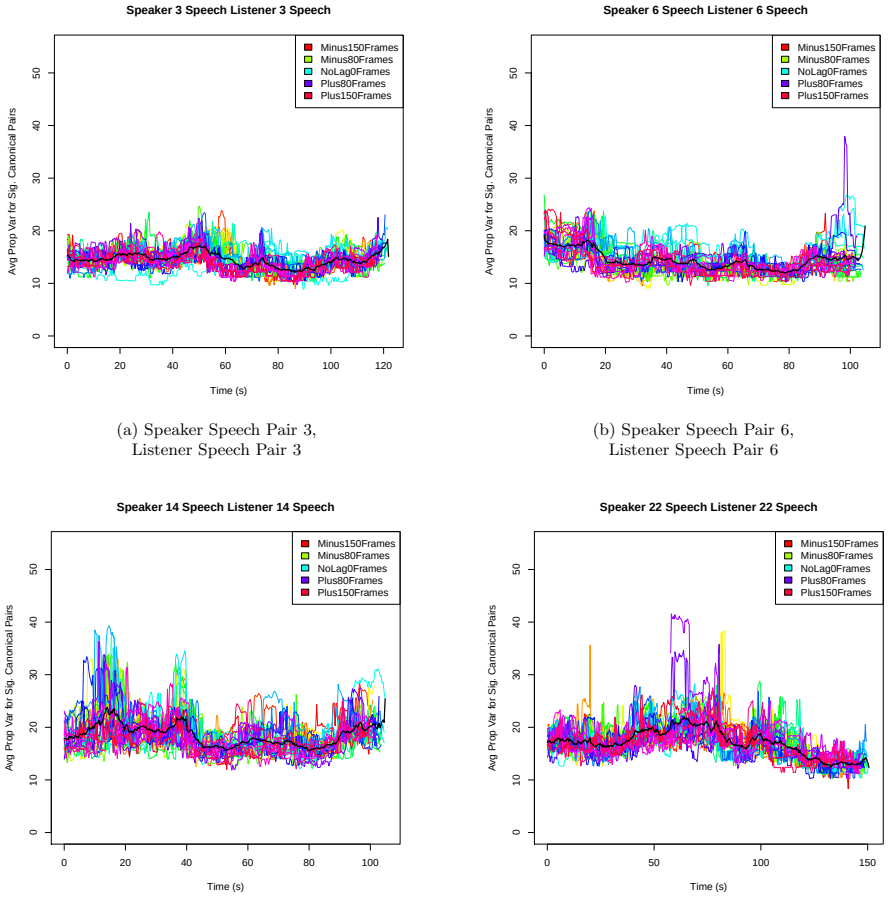



(c) Speaker Face Pair 14, Listener Speech Pair 14




(d) Speaker Face Pair 22, Listener Speech Pair 22

*Figure* A.27: Plots of the average proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's speech for four representative pairs when using the left singular matrices. Note that surrogates for the Face-Speech CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.




(a) Speaker Speech Pair 3, Listener Speech Pair 3

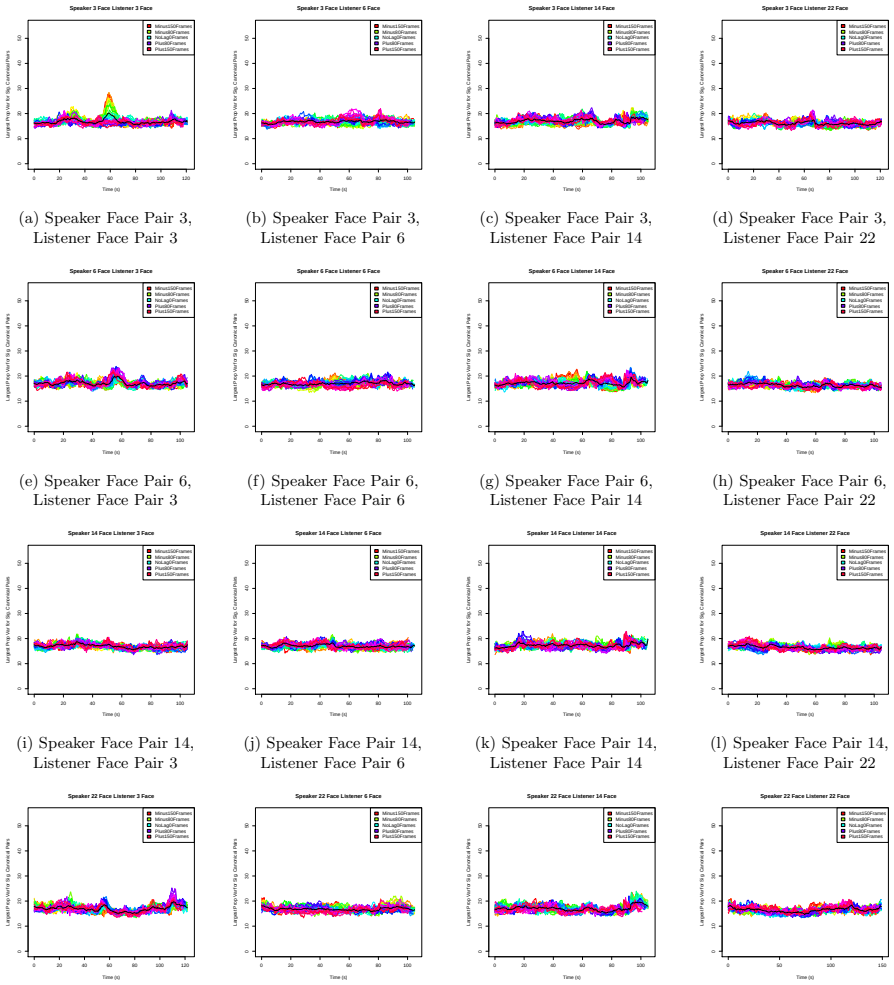
(b) Speaker Speech Pair 6, Listener Speech Pair 6

(c) Speaker Speech Pair 14, Listener Speech Pair 14

(d) Speaker Speech Pair 22, Listener Speech Pair 22

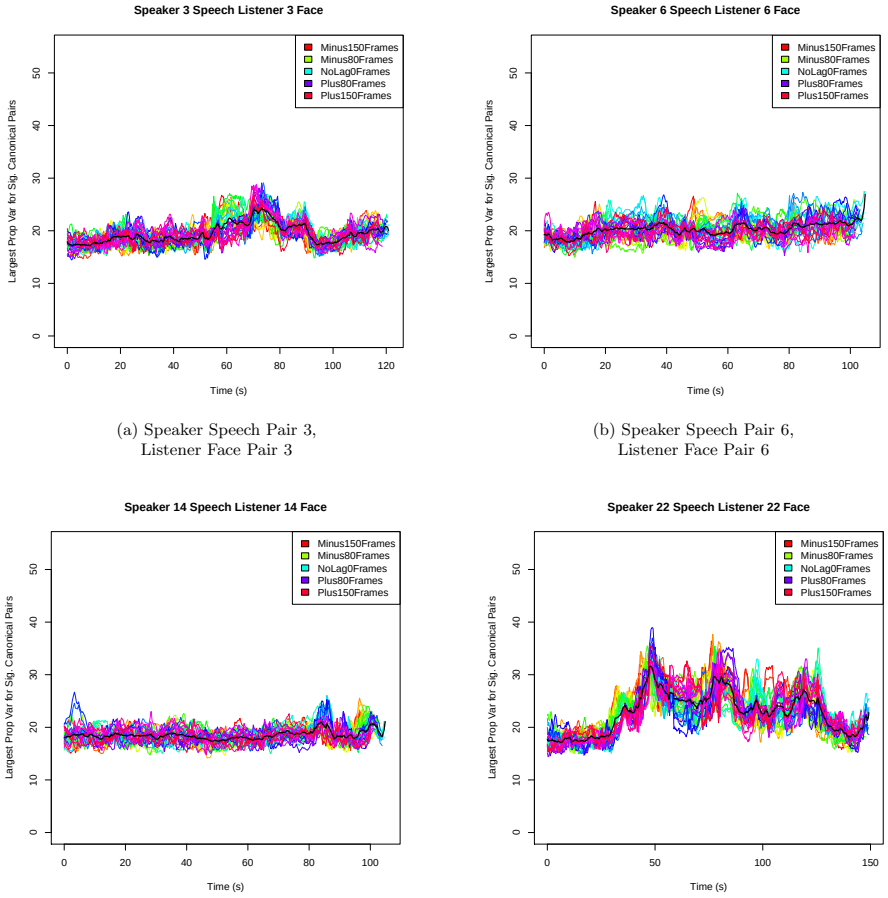
*Figure* A.28: Plots of the average proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's speech for four representative pairs when using the left singular matrices. Note that surrogates for the Speech-Speech CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.

### A.2.4 Largest Proportion Explained



(a) Speaker Face Pair 3, Listener Face Pair 3
(b) Speaker Face Pair 3, Listener Face Pair 6
(c) Speaker Face Pair 3, Listener Face Pair 14
(d) Speaker Face Pair 3, Listener Face Pair 22
(e) Speaker Face Pair 6, Listener Face Pair 3
(f) Speaker Face Pair 6, Listener Face Pair 6
(g) Speaker Face Pair 6, Listener Face Pair 14
(h) Speaker Face Pair 6, Listener Face Pair 22
(i) Speaker Face Pair 14, Listener Face Pair 3
(j) Speaker Face Pair 14, Listener Face Pair 6
(k) Speaker Face Pair 14, Listener Face Pair 14
(l) Speaker Face Pair 14, Listener Face Pair 22
(m) Speaker Face Pair 22, Listener Face Pair 3
(n) Speaker Face Pair 22, Listener Face Pair 6
(o) Speaker Face Pair 22, Listener Face Pair 14
(p) Speaker Face Pair 22, Listener Face Pair 22

*Figure* A.29: Plots of the largest proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's face for four representative pairs and their respective surrogates when using the left singular matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



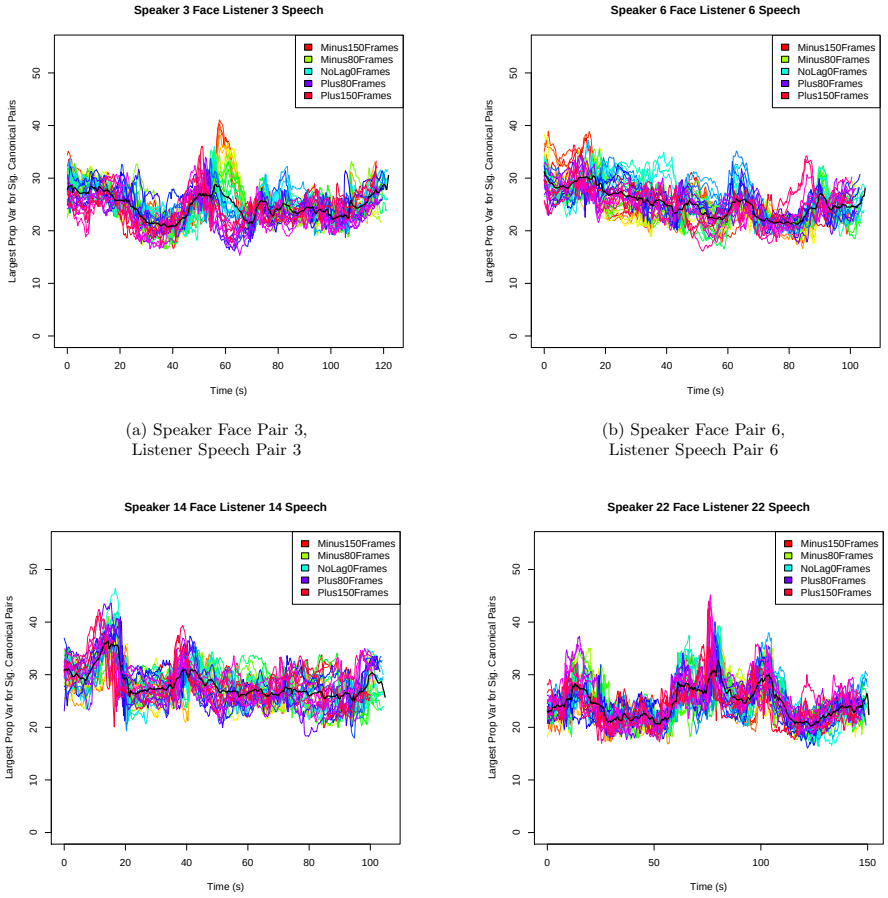

(a) Speaker Speech Pair 3, Listener Face Pair 3

(b) Speaker Speech Pair 6, Listener Face Pair 6

(c) Speaker Speech Pair 14, Listener Face Pair 14

(d) Speaker Speech Pair 22, Listener Face Pair 22

*Figure* A.30: Plots of the largest proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's face for four representative pairs when using the left singular matrices. Note that surrogates for the Speech-Face CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.



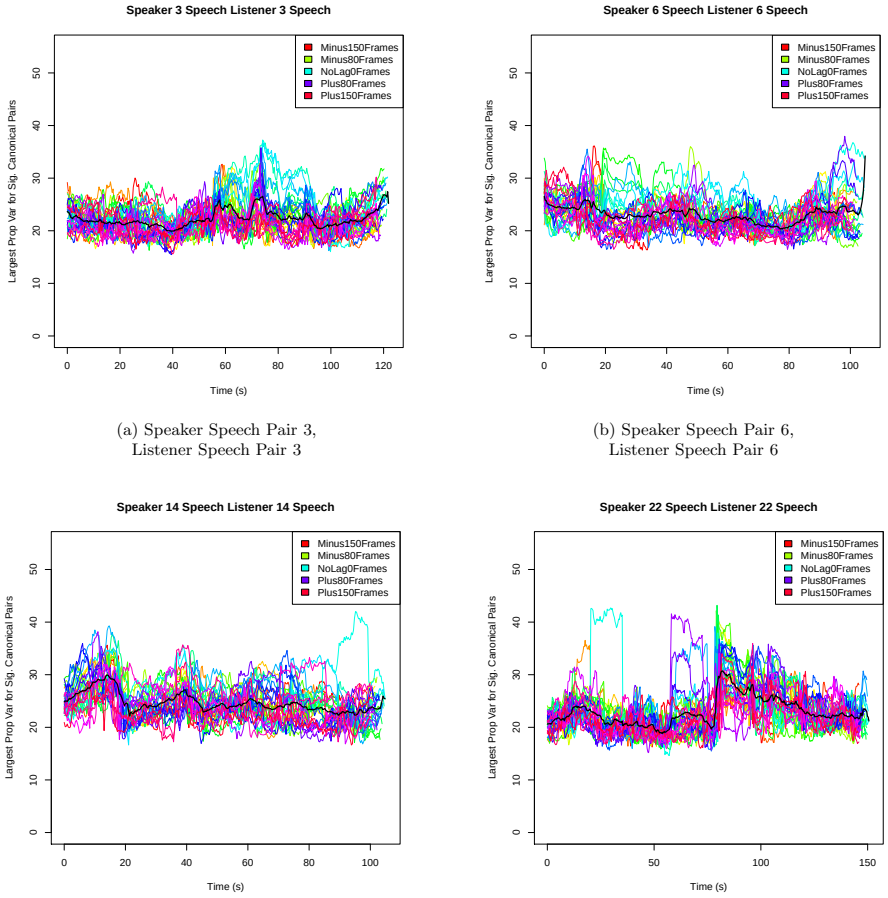

(a) Speaker Face Pair 3, Listener Speech Pair 3

(b) Speaker Face Pair 6, Listener Speech Pair 6

(c) Speaker Face Pair 14, Listener Speech Pair 14

(d) Speaker Face Pair 22, Listener Speech Pair 22

*Figure* A.31: Plots of the largest proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's face and the designated listener's speech for four representative pairs when using the left singular matrices. Note that surrogates for the Face-Speech CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.




(a) Speaker Speech Pair 3, Listener Speech Pair 3

(b) Speaker Speech Pair 6, Listener Speech Pair 6

(c) Speaker Speech Pair 14, Listener Speech Pair 14

(d) Speaker Speech Pair 22, Listener Speech Pair 22

*Figure* A.32: Plots of the largest proportion of variance explained by the significant canonical pairs from the CCA between the designated speaker's speech and the designated listener's speech for four representative pairs when using the left singular matrices. Note that surrogates for the Speech-Speech CCA analyses were only analyzed when using the full matrices. Each colored line represents one lag between the two sets of data, and the solid black line represents the average number of significant canonical pairs for each window.