Conducting Actions Elicit Specific Acoustic Features in How People Vocalize: Cross-modal Correspondence between Gestures and Sounds as a Function of Available Information

By

Aysu Erdemir

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

In partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

August, 2016

Nashville, Tennessee

Approved:

John Rieser, Ph.D.

Amy Needham, Ph.D.

Daniel Ashmead, Ph.D.

Mark Wallace, Ph.D.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor and mentor Prof. John Rieser who has greeted me with enthusiasm to the program several years ago and provided years of intellectual and emotional support. His open-mindedness, never-ending excitement, patience and intelligence have helped me to realize and pursue my own research interests. John, you have not only been a great academic mentor but also a great friend to me and to my family. Words cannot describe my most sincere and profound appreciation for you, I owe you a lot.

I'm forever grateful for Emelyne Bingham, who has been my research partner during this research project. Thank you for all of your valuable contribution to the project, for your incisive comments, your sharp thinking, your help with coding and your never-ending excitement for science of music. It has been beyond a pleasure to work with you.

I am also grateful to each of my other committee members, Dr. Amy Needham, Dr. Daniel Ashmead, and Dr. Mark Wallace for their support, scholarly insights and thought provoking comments, each of which substantially contributed to my dissertation and to my professional development. Special thanks to Amy Needham, who has given me a lot of insightful guidance and feedback for my personal and academic development. Thank you for your smiley face and all the support you have given to me. Further, I am so very grateful for the opportunity to work with Dr. Robin Jones, Dr. Tedra Walden and Dr. Reyna Gordon, each of whom have supported my research interests throughout my training and challenged me to continuously explore the unknown with enthusiasm.

I am grateful to all my lab members I've had the honor to work with during my PhD, whose support and friendship helped me every step of the way. Thanks to all the current and past members of Dynamic Perception and Action Lab, Infant Learning Lab, Developmental Stuttering Lab, and Music Cognition Lab. Thanks to Sara Beck, Jane Hirtle, Sarah Weisen,

Hatun Zengin-Bolatkale, and especially my good friend Maria Osina for their valuable friendship, endless support and comfort during graduate school.

Finally, I would like to thank my family. To my parents Aysel and Suat, thank you for your continual guidance and support, for always trusting that I was making the right decisions, for providing me with such a great foundation, and for loving me. I cannot thank enough for all the support you have given to me thought out all my life and especially during this journey. To my amazing husband, Erdem Erdemir, thank you for your unwavering and infinite support and encouragement. Thank you for listening to me and for understanding me, for cheering me up, and making me laugh, and for giving me the courage. And to my beloved daughter Beste Su, thank you for bringing happiness and joy into my life that I could not ever imagined. Your little presence has helped me to appreciate the beauty of each and every day. I am beyond lucky to have such a great family.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS ii

LIST OF TABLES vii

LIST OF FIGURES viii

Chapter

I. Introduction 1

Perception is multisensory and sensorimotor 2

Action observation - action execution resonance system 3

The correspondence between musical and motional parameters 6

Specific aims and hypotheses 10

Experiment 1. Cross-modal mapping of observed gestures onto vocal sounds 10

Experiment 2. Effect of instruction on gesture sound coupling 11

Experiment 3. Role of musical expertise in gesture sound coupling 11

Experiment 4. Use of point light displays in gesture sound coupling 11

Experiment 5. Role of self-produced auditory feedback in gesture sound coupling 12

Experiment 6. Effect of gestural motor practice in gesture sound coupling 14

II. Method 15

Participants 15

Experimental stimuli 15

Procedure 17

Experiment 1. Cross-modal mapping of observed gestures onto vocal sounds 17

Experiment 2. Effect of instruction on gesture sound coupling 18

Experiment 3. Role of musical expertise in gesture sound coupling 18

Experiment 4. Use of point light displays in gesture sound coupling 18

Experiment 5. Role of self-produced auditory feedback in gesture sound coupling 19

Experiment 6. Effect of gestural motor practice in gesture sound coupling 20

Data analyses methods 21

Kinematic analysis of conducting gestures 21

Analyses of vocal responses 27

Ratings by expert judges 27

Acoustic analyses of vocal responses..........28

Specific hypotheses and statistical analyses methods 29

Hypothesized associations of acoustic and motional parameters..........29

Statistical analyses methods 33

III. Results..........36

Experiment 1. Cross-modal mapping of observed gestures onto vocal sounds 36

Ratings by expert judges 36

Acoustic analyses of vocal responses..........37

Experiment 2. Effect of instruction on gesture sound coupling 39

Ratings by expert judges 40

Acoustic analyses of vocal responses..........41

Experiment 3. Role of musical expertise in gesture sound coupling 43

Ratings by expert judges 44

Acoustic analyses of vocal responses..........45

Experiment 4. Use of point light displays in gesture sound coupling 46

Ratings by expert judges 47

Acoustic analyses of vocal responses..........48

Experiment 5. Role of self-produced auditory feedback in gesture sound coupling 49

Ratings by expert judges 50

Acoustic analyses of vocal responses..........51

Experiment 6. Effect of gestural motor practice in gesture sound coupling 52

Ratings by expert judges 53

Acoustic analyses of vocal responses..........54

Part 1 54

Part 2 62

IV. Discussion..........70

Experiment 1. Cross-modal mapping of observed gestures onto vocal sounds 70

Experiment 2. Effect of instruction on gesture sound coupling 71

Experiment 3. Role of musical expertise in gesture sound coupling 74

Experiment 4. Use of point light displays in gesture sound coupling..........77

Experiment 5. Role of self-produced auditory feedback in gesture sound coupling..........78

General discussion for experiments 1, 2, 3, 4 & 5 81
Origin and shaping of cross-modal correspondences..82
Experiment 6. Effect of gestural motor practice in gesture sound coupling 85
Conclusion...89
Implications 90

REFERENCES 92

LIST OF TABLES

Table Page

1. Laban categorical analysis of the four gestures 17

2. Experimental Procedure for experiment 6 in chronological order..21

3. The acoustic features, what they represent and how they were calculated...............................23

4. Summary of the calculated mathematical values for each of the movement feature................26

5. The acoustic features, what they represent and how they were calculated...............................29

6. Hypothesized cross-modal links 30

7. Categorization accuracies from experiment 1..36

8. The results from the planned comparisons for experiment 1 - nonmusicians 39

9. Categorization accuracies comparing experiment 1 and experiment 2 41

10. The results from the planned comparisons for experiment 2 - weak instruction.....................43

11. Categorization accuracies comparing experiment 1 and experiment 3 44

12. The results from the planned comparisons for experiment 3 – musicians..............................46

13. Categorization accuracies comparing experiment 1 and experiment 4 47

14. The results from the planned comparisons for experiment 4 - point light display 49

15. Categorization accuracies comparing experiment 1 and experiment 5 50

16. The results from the planned comparisons for experiment 5 – masked auditory feedback.....52

17. Percent scores from experiment 6 comparing the motor practice and visual practice.............54

18. a-f. Summary of results from the planned complex comparisons for experiment 6................56

19. a-d. Specific directions of the changes that characterize better performance for experiment 6 along with the results from simple slopes analyses 68

LIST OF FIGURES

Figure Page

1. Representation of the Experimental Stimuli 16

2. Point light display stimuli 19

3. Representation of the gestures being captured in a Motion Capture System 22

4. Representation of the positional data for the four gestures....................24

5. Representation of the velocity data for the four gestures 25

6. Composite velocity patterns superimposed for flick, punch, float, and glide....................26

7. Distribution of categorization accuracies for experiment 1- nonmusicians....................36

8. a-f. Model outputs representing six acoustic features in experiment 1....................38

9. Distribution of categorization accuracies for experiment 1 versus experiment 2....................40

10. a-f. Model outputs representing six acoustic features in experiment 1 versus experiment 2 ..42

11. Distribution of categorization accuracies for experiment 1 versus experiment 3....................44

12. a-f. Model outputs representing six acoustic features in experiment 1 versus experiment 3 ..45

13. Distribution of categorization accuracies for experiment 1 versus experiment 4....................47

14. a-f. Model outputs representing six acoustic features in experiment 1 versus experiment 4 ..48

15. Distribution of categorization accuracies for experiment 1 versus experiment 5....................50

16. a-f. Model outputs representing six acoustic features in experiment 1 versus experiment 5 ..51

17. Distribution of percent accuracies for the motor and visual practice groups 54

18. a-f. The output from the models for each of the six dependent measures in experiment 6 56

19. a-d. Six dependent measures as a function of time (pre-test and post-test) for each of the four gestures 64

CHAPTER I

INTRODUCTION

Music conductors use hand-arm gestures to shape the sound that their musicians produce. The gestures of conductors communicate the information about how the played or sung notes should sound, and music performers seem to understand how to map the features of those gestures onto the features of the sounds they make. The series of experiments presented here is aimed at understanding the nature of such cross-modal links between observed hand gestures and their accompanying vocal responses using several experimental manipulations. The task involved adults watching video clips of four different types of hand gestures (referred to as flicks, punches, floats and glides, following Laban and Lawrence, 1947), and producing the syllable /da/ repeatedly along with the observed gestures. Experiment 1 explored the specific cross-modal links between the kinematic features of the movement and acoustic features of accompanying vocal sounds. Experiment 2 explored the role of instruction and how automatic or deliberate gesture-sound correspondence is. Experiment 3 explored the role of music background, and whether such associations stem from music experience or from everyday life experience. Experiment 4 explored whether perceiving the velocity patterns of the gestures provides sufficient basis for such coupling. Experiment 5 explored the role of auditory feedback and whether gesture-sound coupling is driven auditorily or through the vocal-motor system. And finally, experiment 6, a motor practice study, explored the role of motor representations of the gestures as a potential motor-based mechanism mediating/enhancing this visual-to-auditory mapping.

Perception is multisensory and sensorimotor

Perception is generally a multisensory process (van Atteveldt, Murray, Thut, & Schroeder, 2014). Information for most events around us originates from more than one modality, and most perceptual situations involve sight and sound (and perhaps touch, taste, and smell as well). For example, communication involves both an audible speech and facial/gestural motions or expressions produced in synchrony. Or consider concerts, ballets, and musicals, where we enjoy the abundance of visual and auditory stimulation accompanying one another. In daily life, we are frequently faced with situations where we need to process the dynamic visual and audible information from walking people, driving cars, talking friends, and fuse these inputs into a single coherent representation. The perceptual system has the remarkable ability to fuse sensory input from different modalities into a coherent representation of the world around us, providing us with the opportunity to form countless cross-modal couplings. The fact that many events around us are processed as multisensory events suggests that we might have developed implicit associations between features across modalities as a result of constant exposure to such specific multimodal occurrences (e.g. sight of a heavy object and its association with a low pitch tone).

Because most sensory processing is active, and largely directed by motoric and attentional sampling practices, perception is also sensorimotor in nature (Schroeder et al., 2010). This is particularly clear when sensory events are the direct result of the motor activity. The idea that motor representations are often embedded in the perceptual processes, and that perceptual effects are embedded in motor actions is well established by various researchers (James, 1890; Sperry, 1952; Gibson, 1966; Prinz, 1997; Hommel et. al., 2001; Aschersleben, 2002). According Hommel et al. (2001) perceptual codes (perceived or anticipated events) and action codes (intended or generated events) can influence and prime each other on the basis of their overlap in the representational domain. That is, the activation of either a percept or motor code automatically leads to the activation of the other code, facilitating future execution or perception.

Action observation - action execution resonance system

This idea is central to the psychophysical and neural findings about the mirror neuron system (e.g., Gallese et al., 1996; Rizzolatti & Craighero, 2004). Brain imaging studies with humans show that visual and auditory properties of actions (mere observation of a biological action, or hearing the sounds associated with that action) would automatically activate parts of the motor network that would be used to execute the action itself (Iacoboni et al., 1999; Kilner et al., 2003; Clark et al., 2003; Kohler et al., 2002; Lahav et al., 2007; Ticini et al., 2011). For instance, MEPs recorded from arm muscles are increased when people simply observe basic grasping and arm movements suggesting that there is an action observation and execution matching system (Fadiga et al. 2005).

In line with these ideas, the concept of 'action simulation' provided a possible cognitive framework for the Mirror Neuron System (Jeannerod, 2001). According to this view, we make sense of others' actions by using our own internal motor system as an emulator, i.e. through running an internal neural simulation of the observed action along with its all sensory expectations; in other words, by directly mapping the visual representation of observed action onto our own motor schema. This view provides a possible neural explanation for how people could access the corresponding action knowledge from mere visual perception of an action. Indeed, Bosbach et al. (2005) tested two patients who lacked afferent feedback for touch and proprioception. Although they could perform actions well, they performed poorly when asked to interpret the actions of others, suggesting the lack of actual peripheral sensations resulted in an inability to map the perceptual representation of observed actions onto a representation of the motor pattern for the same action.

Mirror Neurons were discovered and initially studied in the context of basic object-directed actions such as grasping, placing and manipulating objects. Since then, more studies have been conducted to test the idea whether there is a basic physical "resonance" mechanism that maps the kinematic description of observed actions onto one's own motor representations, regardless of the context and goal in which the observed actions are executed. Specifically, cortical excitability and motor resonance have also been documented for "intransitive" movements (movements without use of an object and the motivation of an explicit goal), such as simple arm flexion movements (Melzoff & Prinz, 2002), thumb ab-/adduction (Maeda et al., 2002), tracing geometrical shapes in air, or simple arm lifting (Fadiga et al., 1995), and communicative hand gestures that do not involve an object (Montgomery et al., 2007). These studies have confirmed that muscles that are activated during the observation of a given action are the same as those activated when the action is physically executed. Moreover, Gangitano et al. (2001) and Montagna et al. (2005) showed a strict temporal coupling between the kinematics of an observed reaching-grasping movement and the modulation of the cortical motor excitability in the observer. Maeda et al. (2002) further provided evidence about the degree of specificity in the motor cortical excitability induced by observation of finger movements. They have shown that the degree of modulation is maximal when the observed hand orientation corresponds with that of the observer (hand presented on the screen was facing out from and corresponding to that of the observer). In addition Aziz-Zadeh et al. (2002) evidenced that motor facilitation is lateralized such that observing right hand movements excite left motor cortex and vice versa.

In summary, physiological studies have shown that a resonant motor mapping mechanism, is present not only during goal-directed actions but also during intransitive movements; is specific for the muscles involved in the observed movements; and is even temporally coupled with the kinematics of observed action. This resonance mechanism, which can also code for meaningless movement segments that form actions, can be used for important functions, such as imitation learning which requires a direct matching of observed actions onto existing motor codes.

Buccino et al. (2004) have shown that there is a basic resonance circuit involving inferior parietal lobe, inferior frontal gyrus and premotor cortex underlying imitation learning; and these regions start to get activated during passive action observation with additional areas partaking during active imitation. Using TMS, Clark et al. (2003) recorded MEPs form the dominant hand while participants watched simple videotaped right hand movements, and they looked at cortical excitability induced by *passive observation* and compared this with activity induced by *active imitation*. Passively observing another person's action resulted in the activation of the same basic motor pathways as actively imitating the action, although it was to a lesser degree than in active imitation. Moreover, Viviani and Stucchi (1989, 1992) showed that when perceptual velocity of drawing and writing trajectories violated the two-thirds power law [which is the motor rule for human motion when drawing curved trajectories], the perceptual judgments were inaccurate. This suggests that people go beyond the information provided by visual stimulus and infer the hidden kinematics of drawing movements behind the visual information by using their own motor expertise.

Similarly, Flanagan and Johansson (2003) asked people to observe a manual block stacking task. They found that people did not just passively observe on a perceptual level, and instead they automatically coordinated their gazes with the stimuli in a predictive way, as if performing the action themselves. Su and Jonikaitis (2011) brought action and perception overlap even one step forward by showing that embedded motor representations underlying a visual motion and an auditory motion (tempo) can interact across modalities to form a coherent experience. Specifically, they showed that observed visual movements are automatically transformed into action patterns, which then biases the perception of auditory tempo. These studies supported the view that action knowledge is implicitly involved during perceptual processing of movements; and that information from multiple modalities can affect one another through a motor code underlying the perceptual events.

The correspondence between musical and motional parameters

Musical contexts serve as an ideal case for studying multi-modal sensory-motor interactions because each action in a music performance (whether observed or self-executed) is intended to produce concurrent sounds. Gestures of orchestra conductors, for instance, involve complex dynamic spatio-kinetic cues, which need to be conveyed to fellow musicians through vision. An ensemble/choir member following a conductor's lead should be able to perceive and encode the dynamic visual information present in conductor's gestures, and translate this into appropriate auditory responses through use of his own motor system, which seems to portray highly complex multisensory–motor interplay. Such contexts serve as an ideal case for understanding how auditory and visual modalities are cross-matched, and the role of motor system within this coupling.

It is an old and common assumption that music and movement are closely related. For instance, until the late 19th century, before the invention of sound recording, music performances were always produced and experienced as movement-sound integrated activities. Moreover, there is a rich terminology used by musicians that refer to music in motional terms (e.g. lento-slowly, corrente-running, andante-walking, con moto-with movement). And, above all, it has been known that people have a natural tendency to react to music in motional terms, by tapping, rocking, singing, or dancing. Moving rhythmically to music has been observed in all known cultures (Brown, 2003). This effect has been well documented even in very young infants (Zentner & Eerola, 2010; Philips-Silver & Trainor, 2008), as well as some bird species such as cockatoo (Patel et al, 2009), parrot (Schachner, Brady, Pepperberg, & Hauser, 2009), and budgerigar (Hasegawa, Okanoya, Hasegawa, & Seki, 2011), as well as a sea lion (Cook, Rouse, Wilson, & Reichmuth, 2013). Moreover, several neuroimaging studies reported the tight coupling between the motor and auditory systems by showing that specific motor areas, in addition to being instrumental in movement production, mediate music perception (especially rhythm and beat) as well (Chen, Penhune and Zatorre, 2008; Grahn and Brett, 2007; Alluri et al., 2012). Additionally, computational kinematic models have been developed to bring out the tight coupling between sound and motion by offering relationships between the laws of physical action in the real world and expressive timing in music (Sundberg & Verillo, 1980; Todd, 1992; Friberg & Sundberg, 1999; Erdemir et al., 2010).

Other studies show that perceptual experience of music performance is an inherently audio-visual phenomenon, emphasizing the role and importance of vision for a complete musical experience. These studies often emphasize the importance of physical movement for musical communication by showing that audiences can perceive expressive (Davidson, 1993), emotional (Dittrich et al., 1996), and structural characteristics (Schutz & Lipscomb, 2007) of music through vision of movement alone. For instance, by studying marimba players, Schutz & Lipscomb (2007) have shown that notes of certain duration are perceived to be longer when paired with a long gesture than when paired with a short gesture, suggesting vision of action often biases auditory perception.

Empirical studies investigating auditory and motional correlations mostly have looked at the influence of musical parameters on body movements, by observing people when they actively engage in an action in response to auditory stimulations. In a study by Eitan and Granot (2006) participants listened to auditory stimuli consisted of several manipulations in dynamics, pitch contour, pitch intervals and articulation, and they were asked to imagine a human cartoon character moving in accompany with the melodic patterns. They gave their responses based on a set of pre-determined motion dimensions. In a relevant study by Kohn and Eitan (2009) children were asked to move their bodies to various musical excerpts from classical music that involved changes in pitch, loudness and tempo. The children were simply asked to move in a way that would match the musical patterns they were listening to. Similarly, Küssner et al (2013) asked adults to visually represent pure tone sequences varying in pitch, loudness and tempo on a tablet by drawing with a pen which connected to a pressure sensor. Nymoen, Godoy, Jensenius and Torresen (2012) asked participants to move their hands while listening to various short sound objects that varied in pitch, spectral centroid and intensity. And finally, Burger et al. (2011) asked adults to move in a way that feels natural to various musical excerpts from different musical genres ranging from pop, jazz to hip hop. Several computational and motion capture techniques have been used to look at the relationship between movement kinematics and musical features, and investigated musicians' bodily movements (Thompson & Luck, 2008), dance movements (Burger et al., 2011), and conductors' gestures (Luck & Toiviainen, 2006).

Another line of research focused on perceptions of congruence between simple auditory-visual (AV) combinations in musical contexts. For instance, Lipscomb and Kim (2004) asked participants to rate the match between various AV components which include manipulations in basic acoustic features (such as pitch, loudness, timbre, duration), and visual features (such as color, shape, size, verticality). Similarly, Kohn and Eitan (2012) asked participants to rate the match between various videotaped dance moves with simple changes in pitch and loudness.

All of these studies have found various systematic variations in the motion features based on changes in acoustic features in the musical samples. Simply, changes in tempo, pitch and loudness have been found to be associated with changes in speed, verticality and energy. These studies also generally found only few differences between musically trained and untrained participants (Eitan & Granot, 2006; Küssner, Gold, & Leech-Wilkinson, 2012; Küssner, Tidhar, Prior, & Leech-Wilkinson, 2014). When the differences existed, they did not indicate opposite tendencies, but rather stronger and more consistent tendencies by musicians to associate particular auditory and motional changes. This suggested that motion-sound associations generally stem not from musical connotations but from more general non-musical sources.

Past research investigating associations of acoustic changes with visible motion has mainly focused on either perceptions of congruence between simple auditory-visual combinations (Lipscomb and Kim, 2004; Kohn and Eitan, 2012), or at the influence of musical parameters on body movements (Eitan & Granot 2006; Kohn & Eitan, 2009; Küssner et al., 2012, Küssner et al., 2014; Thompson & Luck, 2008; Burger et al., 2011). Their methods involved a purely psychoacoustic perspective that focuses on bringing out the specific audio-visual associations, with no intention for exposing the underlying cognitive mechanism, nor did they discuss a possible role of the motor system for the observed cross-modal links. Moreover, they exclusively used highly controlled musical segments as the auditory stimuli.

No study, to our knowledge, has investigated changes in acoustic parameters of spoken utterances based on ecologically valid observed gestural movements, which closely mimics the situation between a conductor and a vocal soloist. Moreover, no study, to our knowledge, investigated the nature of such relationships by systematically manipulating the kind of information available to the participants as they engage in a visual-to-auditory mapping task. Past research has suggested that these cross-modal associations mostly arise from everyday life experience as they are present even in musically naïve participants, but we do not know whether the strength of such relationships depend on factors such as the instruction given to the participants, the kinematic information available in the visual gesture, the absence or presence of self-produced auditory feedback, and practice with of motor representations underlying observed gestures.

Therefore, the current research aims to combine the two lines of research from motor cognition and music cognition by investigating the role of instruction, musical experience, spatio-kinematic cues, auditory feedback and motor representations in the cross-modal mapping of movement to sound. In the first experiment we will specifically investigate whether there are systematic variations in the vocal responses based on the observed gestures through use of detailed acoustic and movement analysis. Subsequent series of studies will further investigate the underlying nature of such cross-modal correspondence by systematically degrading/enhancing the kind of sensory/motor information available to the participants in various ways; and by looking at how and in what ways these manipulations affect the strength of the specific motional and auditory links.

Specific Aims and Hypotheses

Experiment 1. Cross-modal mapping of observed gestures onto vocal sounds

Aim: is to explore whether there is a systematic relationship between four different hand gestures performed by an expert conductor, and accompanying vocal sounds produced by college students with no significant amount of music background.

Hypothesis: Participants will systematically vary their spoken utterances in a way to match the motion characteristics of the visually observed gestures. The acoustic parameters of the spoken /da/ sounds [e.g. duration, amplitude, amplitude variability, fundamental frequency, pitch variability and vowel quality] will be associated with the movement features of the conductor’s gestures [e.g. duration, velocity, spatial displacement].

Experiment 2. Effect of instruction on gesture sound coupling

Aim: is to explore whether such coupling is based on an automatic processing/mapping of visual features onto acoustic features, or on a deliberate cognitive strategy. The instruction given to the participants was minimized so that no hint was given about the desired match between gestures and sounds.

Hypothesis: Gesture-sound coupling is not entirely deliberate/strategic. We have predicted that the participants would vary at least some of the basic acoustic features unintentionally even with minimal instruction.

Experiment 3. Role of musical expertise in gesture sound coupling

Aim: is to explore whether musical expertise strengthens gesture sound mapping. If musicians relate acoustic and motional features in similar ways, it would support the hypothesis that cross modal links do not stem from musical experience but instead stem from more general sources.

Background: Previous literature generally found only few differences between musically trained and untrained participants (Eitan & Granot, 2006; Küssner, Gold, & Leech-Wilkinson, 2012; Küssner, Tidhar, Prior, & Leech-Wilkinson, 2014), and when differences existed, they indicated stronger and more consistent tendencies by musicians.

Hypothesis: Based on the literature, we have predicted that musicians and non-musicians will vary the acoustic features of their responses similarly, but specific associations will be strengthened as a function of musical practice.

Experiment 4. Use of point light displays in gesture sound coupling

Aim: is to explore whether gesture-sound links remain as strong when the featural body information are eliminated from the visual stimuli and they contained spatio-kinematic information only. This was achieved by presenting the participants with motion-capture based dynamic point light patterns (PLD) of the four gestures.

Background: Past research has shown that point light displays (originally described by Johansson, 1973) are sufficient to convey e.g. gait of friends (Dittrich, Churchill, & Weidenbacher, 1994), gender of a walking person (Kozlowski & Cutting 1977; Mather & Murdoch 1994), the type of human action (Dittrich, 1993) and basic emotions portrayed by body movements (Atkinson, Dittrich, Gemmel and Young, 2004; Brownlow, Doxin, & Radcliffe, 1997; Dittrich, Troscianko, & Morgan, 1996). However, to our knowledge, there is no study that looked at whether full body representation of gestures is necessary to reliably map movement characteristics onto sounds.

Hypothesis: Based on the previous research summarized above we have predicted little/no change in the performance when the videos viewed involved motional information (positional and velocity information) only.

Experiment 5. Role of self-produced auditory feedback in gesture sound coupling

Aim: is to explore the role of self-produced auditory feedback during cross-modal mapping of gestures and sounds. This was achieved by masking the auditory feedback available to the participants while they are vocalizing to match the gestures they observed. We manipulated the availability of self-produced auditory feedback in order to probe how people match gestures with vocalization when they were deprived of auditory feedback and needed to rely solely on motor information and kinaesthetic feedback. Specifically it is aimed at exploring whether such mapping is a visual-to-auditory mapping, or visual-to-vocal-motor mapping.

Background: Vocal performance is known to depend on the skilful use of auditory feedback (Raphael, Borden, & Harris, 2007). Clinical and experimental studies show that audition plays an important role in vocal control during infancy, childhood, and adulthood. For example, past research has shown that postlingual profoundly deaf adults have higher levels of fundamental frequency (F0) (Leder, Spitzer, & Kirchner, 1987a), increased variations in F0 (Lane & Webster, 1991), and difficulties controlling vocal intensity (Leder, Spitzer, Milner, Flevaris-Phillips, Kirchner, & Richardson, 1987c) as well as speaking rate (Lane & Webster, 1991, Leder, Spitzer, Kirchner, Flevaris-Phillips, C., Milner, P., & Richardson, 1987b; Plant, 1984). Moreover, other studies have demonstrated that lack of auditory feedback leads to deterioration of intonation accuracy and fine control of F0 while singing (Elliot & Niemoeller, 1970; SchultzCoulon, 1978; Murbe, Pabst, Hofmann, & Sundberg, 2002; Ward & Burns, 1978). Erdemir & Rieser (2016) have shown that trained singers rely less on auditory feedback for accurate control of pitch when singing a well-known song, compared to instrumentalists and nonmusicians. However, no study, to our knowledge, has investigated the role played by auditory feedback and motor representations in a visual-to auditory cross modal task.

Hypothesis: If the coupling is disrupted when auditory feedback is absent, it would suggest that people rely on hearing their auditory output for such mapping, and that it is an example of visual to auditory matching. If, on the other hand, the coupling is not disrupted under the absence of auditory feedback, it would suggest that people rely on their vocal motor representations and kinaesthetic sensations for such mapping, and that they map the visual information into the vocal-motor system directly without the intervening effect of auditory feedback. Second option would also suggest that vocal motor system is enough for reliably mapping visual gestures onto acoustic features.

Experiment 6. Effect of gestural motor practice in gesture sound coupling

Aim: The last specific aim was to probe the kind of underlying mechanism mediating the coupling of visual gestures and acoustic sounds, and to explore whether active motor practice of observed gestures enhances the visual-to-auditory coupling.

Hypothesis: The hypothesis is that the coupling is (at least partially) mediated by the motor system, and that during visual observation of the gestures the motor representations underlying the observed gestures activate the movements needed to produce and modulate the spoken /da/ sounds. In that respect, we predicted that overt motor imitation (compared to passive viewing) of the observed gestures would lead to increased cortical excitability, which would help to strengthen visual-to-auditory mapping and result in tighter coupling of the observed gestures and spoken sounds. This would also suggest that the structural similarity between observed movements and vocal responses is enhanced when the motor representations underlying the execution of the gestures are incorporated into one's own motor repertoire. If overtly activating the motor pathways corresponding to the execution of observed gestures helps to strengthen the match between gesture and sound, it would suggest that motor representations underlying observed sequences play a role in the cross-modal transfer from vision to sound.

CHAPTER II

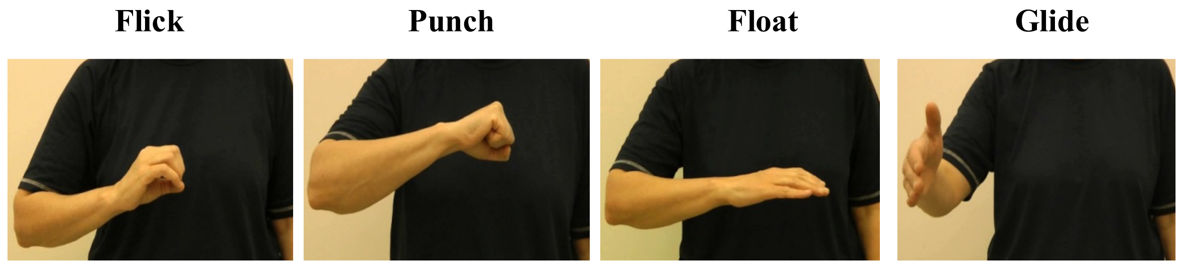
METHOD

Participants

The participants for experiments 1, 2, 4 and 5 were a different set of 20 adults (14F, 6M) recruited from the student body of Vanderbilt University with no significant music background. For experiment 3, the participants were 20 young musicians (14 F, 6 M) who were recruited from the student body of Blair School of Music at Vanderbilt University, with a background of formal music lessons of at least 8 years. For experiment 6, the participants were 48 adults (18 F, 6 M in each group) recruited from the student body of Vanderbilt University with no significant music background.

Experimental stimuli

The participants were asked to view the video recordings of a professional conductor performing four different right hand gestures, called flicks, punches, floats and glides, 10 times in a row at a constant tempo (60 bpm), and to utter /da/ sounds in a way that would match what they observe visually. The video clips involve 10 repetitions of the same gesture, which takes 10 seconds in total. In an effort to isolate the hand gestures and remove facial emotional cues, the videos show only the conductor's chest, shoulders, arms and hands in view as shown in Figure 1.



Flick

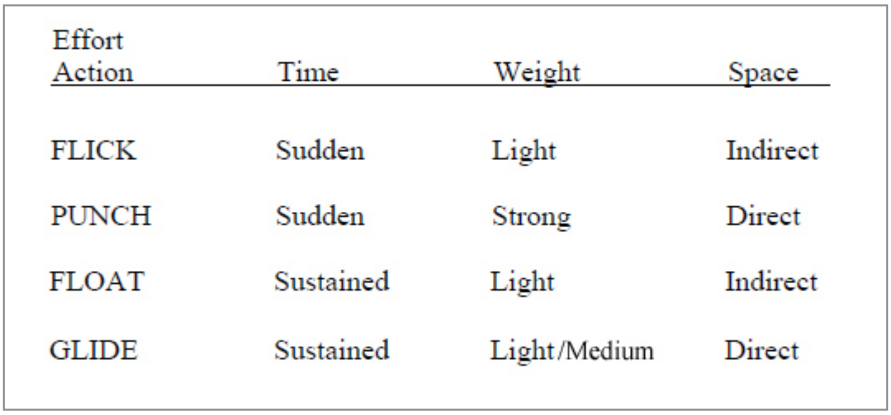
Punch

Float

Glide

Figure 1. Representation of the Experimental Stimuli

The conducting gestures are part of the Effort Actions proposed by Laban and Lawrence (1947). These gestures were selected purposively, since on the one hand they are not contained in conducting manuals, but on the other hand they are meaningful, familiar and natural in an everyday life context. The gestures varied in terms of their use of *time* (sudden/sustained), *weight* (strong/light) and *space* (direct/indirect), as shown in Table 1. By use of time Laban meant the time the act needs to be completed. Sudden gestures are urgent, quick and hasty. Sustained gestures are taking time, leisurely. By use of weight Laban meant the force effort, qualitative use of energy. Light gestures are delicate and airy; and strong gestures are impactful, vigorous and powerful. By use of space Laban meant the manner the space is approached. Direct gestures are channeled, following a fix line, and linear. Indirect gestures are flexible, roundabout and scanning. Laban also reported every day analogies for each gesture. Flick represents removing an insect from a dress; punch represents across and downward hit as in boxing; float represents cradling a soap bubble; and glide represents using an iron to smooth out materials (Harlan, n.d.).



Effort Action	Time	Weight	Space
FLICK	Sudden	Light	Indirect
PUNCH	Sudden	Strong	Direct
FLOAT	Sustained	Light	Indirect
GLIDE	Sustained	Light/Medium	Direct

Table 1. Laban categorical analysis of the four gestures. Categorical analysis of the movement characteristics for each of the four gestures according to Laban Movement Analysis (LMA). LMA is a method widely used in various disciplines, including dance, drama, physical therapy, as well as behaviour research in psychology, anthropology, and other fields (Laban and Lawrence, 1947). Table from Harlan (n.d.)

Procedure

Experiment 1. Cross-modal mapping of observed gestures onto vocal sounds

The participants were positioned in front of a monitor connected to a Dell laptop computer, from which the video presentation was controlled. The participants wore a headset microphone (Audio Technica) positioned approximately 2 inches away from the mouth, and they were instructed to keep the microphone at the same location throughout the whole procedure. Vocal samples were digitally recorded to a Dell laptop computer by means of a digital audio editor program (Adobe Audition 3.0) for later analysis.

The procedure began with participants viewing the four distinct gestures in random order, which is the familiarization phase. In this phase, the participants were asked to "silently watch the four videos, while paying attention to how the gestures differ." In the following (test) phase they were asked to produce the syllable /da/ out loud repeatedly in a way that feels natural along with the gestures they viewed visually. The test phase took place in 2 blocks of four gestures being presented successively and in random order. The specific instruction was as follows: "Please say the syllable /da/ in a way that you think would naturally “match” the gesture you observe. There is no right or wrong way to do it”. No other instruction was given. The task was defined as speaking /da/ instead of singing a melody, first because most of our participants lacked musical background and they would possibly pay more attention to their intonation than to the task itself if they were asked to sing; and second because acoustical analyses could be performed more reliably on spoken sounds than on sung melodies of different pitches where different sung frequencies could act as a confounding variable. All subsequent experiments used the same experimental procedure with specific manipulations described below:

Experiment 2. Effect of instruction on gesture sound coupling

In this experiment we manipulated the strength of the instruction by removing the key word “match” from the instruction. The “weak” instruction given to the participants was “Please say the syllable /da/ *along with* the gestures you view. There is no right or wrong way to do it”. Participants completed 2 blocks with the weak instruction.

Experiment 3. Role of musical expertise in gesture sound coupling

In this experiment we manipulated prior music experience, by recruiting music majors with at least 8 years of formal music training. Participants completed 2 blocks.

Experiment 4. Use of point light displays in gesture sound coupling

In this experiment we manipulated the visual stimuli, by creating motion-capture based point light displays (PLD) of the gestures, which included only kinematic cues of spatial extents, velocities and accelerations. We have recorded the movements of the conductor performing the four gestures in a motion-capture system (Vicon) with reflective markers attached to the joints of the upper torso, and then created movie clips out of these recordings to be presented to the participants. The participants did not have any trouble identifying movements from the stick figures, and they completed 2 blocks viewing the point light displays. Snapshots of the dynamic stimuli appear in Figure 2

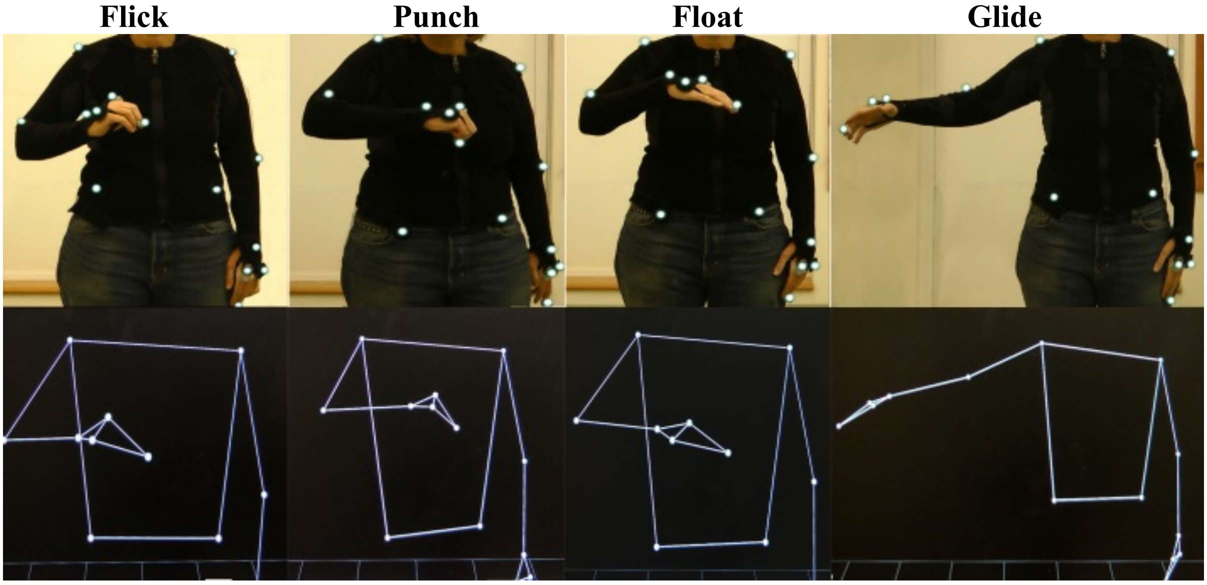


Figure 2. Point light display stimuli. Upper panel represents the recording session, and the lower panel represents the point light displays presented to the participants.

Experiment 5. Role of self-produced auditory feedback in gesture sound coupling

In this experiment we manipulated the availability of self-produced auditory feedback by masking auditory feedback using a babble mask so the participants needed to rely solely on motor information and kinaesthetic feedback for matching gesture onto sound. The experimental manipulation involved the same paradigm used in Erdemir & Rieser (2016). To exclude air and bone-conducted hearing, participants wore a set of in-ear passive sound isolating earphones (Creative MZ0365 EP-830) through which they heard a masking stimulus (Babble-mask).The babble-mask consisted of the sound of 20 adults talking simultaneously, which is completely unintelligible to the listeners. Babble-mask has been shown to be an effective auditory mask by Jones & Keough (2008) and Keough at al. (2013). The optimum volume level of the masking stimulus necessary for accurate masking was identified on an individual basis. In order to achieve this, the volume of the masking noise was gradually increased while participants produced the syllable /da/ repeatedly until they reported they could not hear their own voices any more. After the task each participant was specifically instructed to verify that that they did not hear themselves with the masking stimuli. The participants completed the 2 blocks without being able to hear themselves.

Experiment 6. Effect of gestural motor practice in gesture sound coupling

Experiment 6 builds on experiment 1 by adding a motor practice intervention and a visual practice intervention between pre-test (2 blocks) and post-test (2 blocks). The purpose of the manipulation was to evaluate the possible benefits of a brief motor practice task on the ability to map visual gestures onto accompanying sounds. In the pre-test all participants completed 2 blocks of the gesture-sound matching paradigm, followed by an intervention (motor or visual), and then after the intervention they all completed another set of gesture-sound matching paradigm (post-test). Half of the participants engaged in a motor practice task and the other half engaged in a visual practice task. The visual practice acted as a control task for the possible benefits of motor practice. The motor practice group (n=24) actively imitated the gestures while the visual practice group (n=24) silently viewed each gesture, for a total of 8 times (80 minutes). During the motor practice the participants were instructed to watch the video clips as they actively imitate each gesture with their dominant hand/arm as accurately as possible without producing any sounds. The specific instruction was as follows: "please imitate the gesture as accurately as possible, and please try not to rehearse the /da/ task in your head as you are doing this". The experimenter observed their gestures and provided verbal feedback if their movements did not correspond with the target gestures. (e.g. "Please try to produce a more abrupt movement", or "please try to move your hand at a more constant speed"). During the visual practice the participants passively and silently viewed each gesture for the same amount of time (80 minutes). The specific instruction was as follows: “please silently watch the gesture as you pay attention to the movement, and please try not to rehearse the /da/ task in your head as you are doing this”. Immediately after each practice, the participants performed the cross modal matching task with the corresponding gesture. A schematic of the experimental procedure appears in Table 2.

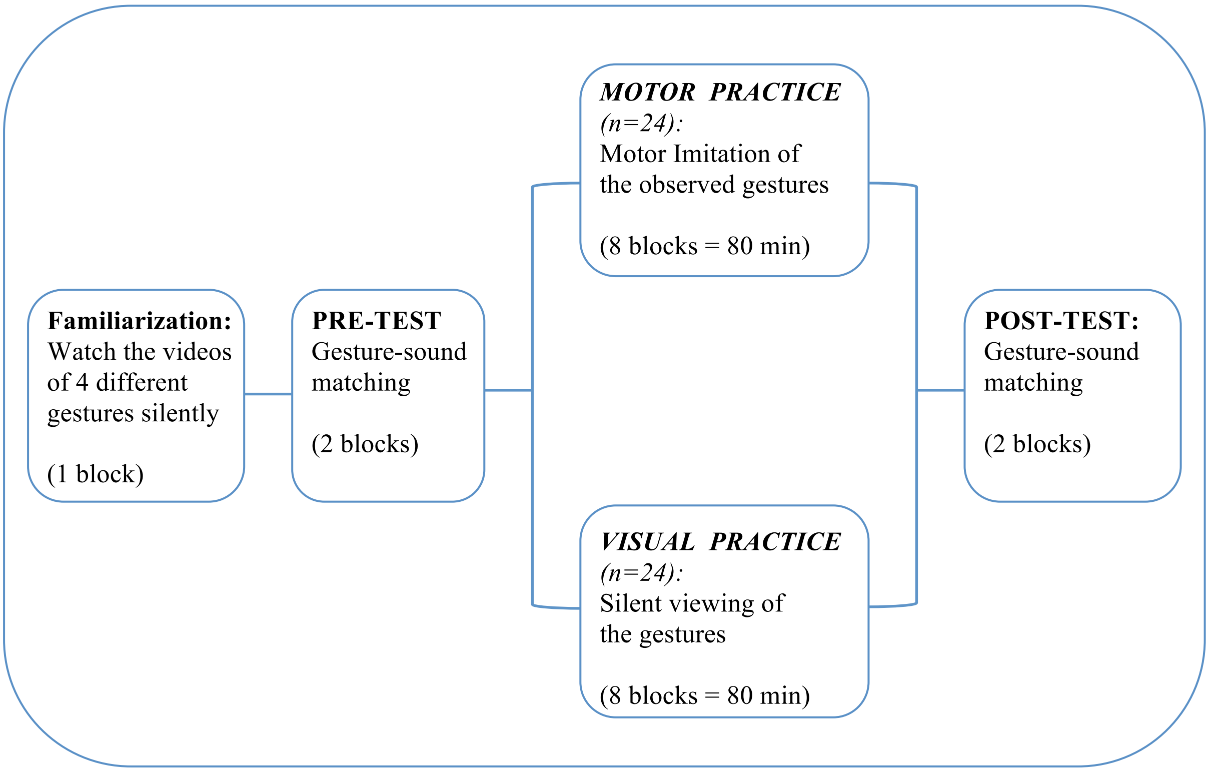

Familiarization: Watch the videos of 4 different gestures silently
(1 block)
PRE-TEST Gesture-sound matching
(2 blocks)
MOTOR PRACTICE *(n=24):* Motor Imitation of the observed gestures
(8 blocks = 80 min)
VISUAL PRACTICE *(n=24):* Silent viewing of the gestures
(8 blocks = 80 min)
POST-TEST: Gesture-sound matching
(2 blocks)

Table 2. Experimental Procedure for Experiment 6 in chronological order.

Data analysis methods

Kinematic analysis of conducting gestures

We have recorded the gestures through reflective markers attached to the joints of the body of the conductor in a motion capture system (Vicon). We have used the data from one marker only- marker attached to the joint of the middle finger combining to the hand, since this marker provided the best representation of the whole movement. The position of the markers and a representation of the four different gestures as being captured appear in Figure 3.

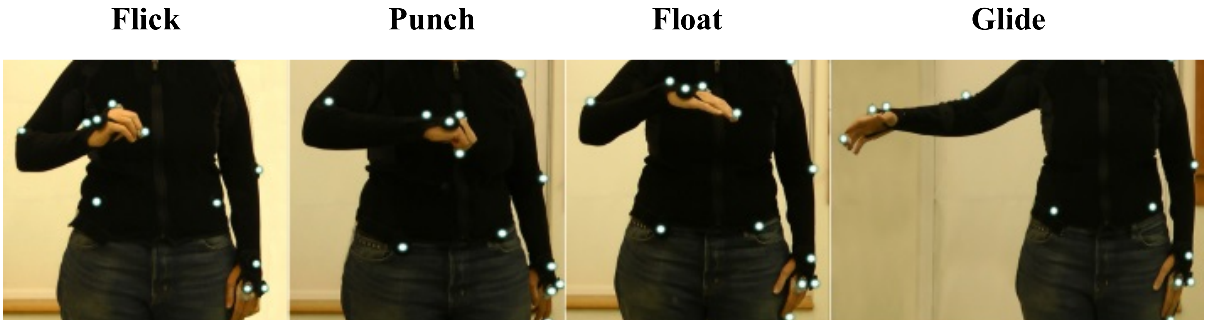

Flick **Punch** **Float** **Glide**

Figure 3. Representation of the gestures being captured in a Motion Capture System.

The gathered position data (x, y, z coordinates) was then imported into Matlab, from which several basic kinematic components are extracted. The movement features that were extracted, how they were computed, and what they represent are summarized in Table 3.

Movement Feature	Description and formula
Duration (ms)	Duration from the beginning until all three (x,y,z) coordinates reaches zero velocity
Horizontal Displacement (HD)	The sum of distance between each consecutive data point in the horizontal position (x-axis). $\sum_{i=1}^{100} \lvert x_{i+1} - x_i \rvert$
Vertical Displacement (VD)	The sum of distance between each consecutive data point in the vertical position (y-axis). $\sum_{i=1}^{100} \lvert y_{i+1} - y_i \rvert$
Sagittal Displacement (SD)	The sum of distance between each consecutive data point in the sagittal position (z-axis). $\sum_{i=1}^{100} \lvert z_{i+1} - z_i \rvert$
Composite Displacement (CD)	Euclidean distance between the three vectors, as calculated by the formula: $\sum_{i=1}^{100} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2}$
Mean Horizontal Velocity (MHV)	The mean value of the first derivate of horizontal position (x-axis). $\frac{dx}{dt}$
Mean Vertical Velocity (MVV)	The mean value of the first derivate of vertical position (y-axis). $\frac{dy}{dt}$
Mean Sagittal Velocity (MSV)	The mean value of the first derivate of sagittal position (z-axis). $\frac{dz}{dt}$
Mean Composite Velocity (MCV)	The mean value of the first derivative of composite displacement.
Composite Velocity Variation (CVV)	The standard deviation of normalized Composite Velocity.
Mean Initial Velocity (MIV)	The mean value of Composite Velocity until the movement reaches peak velocity (first 10 ms out of 100 ms).
Maximum Elevation (ME)	The maximum elevation of the hand from the floor

Table 3. The list of movement features extracted from motion capture. Table summarizes what they represent and how they were computed.

The horizontal, vertical and sagittal position profiles for each gesture appear in Figure 4. The horizontal, vertical and sagittal velocity profiles for each gesture appear in Figure 5. Combined velocity profiles superimposed for each gesture appear in Figure 6. Table 4 summarizes the calculated mathematical values for each of the movement feature extracted for each gesture.

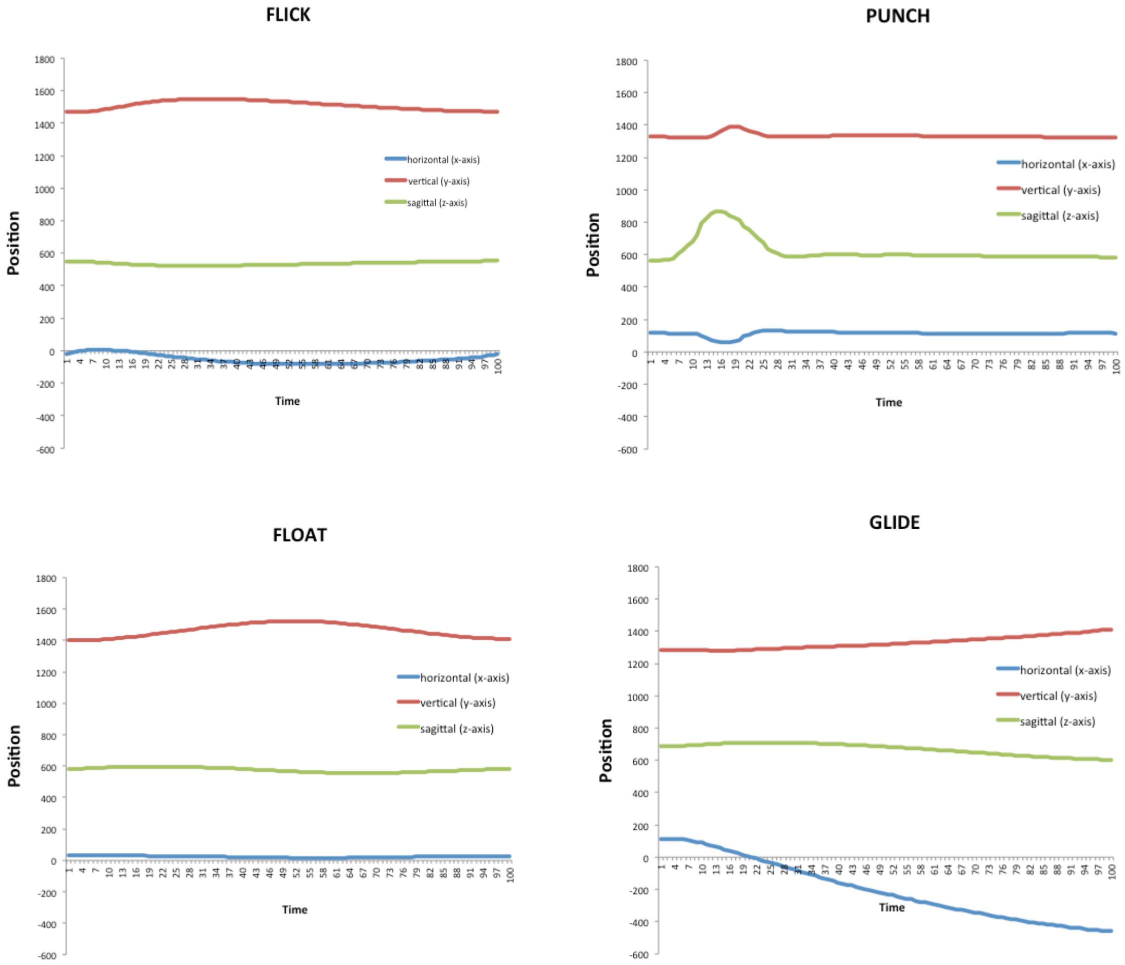


Figure 4. Horizontal (x-axis), vertical (y-axis) and sagittal (z-axis) position data for flick, punch, float and glide. The graphs represent position (mm) on the y-axis, and duration on the x-axis (1000ms-1sec) for each of the four gestures.

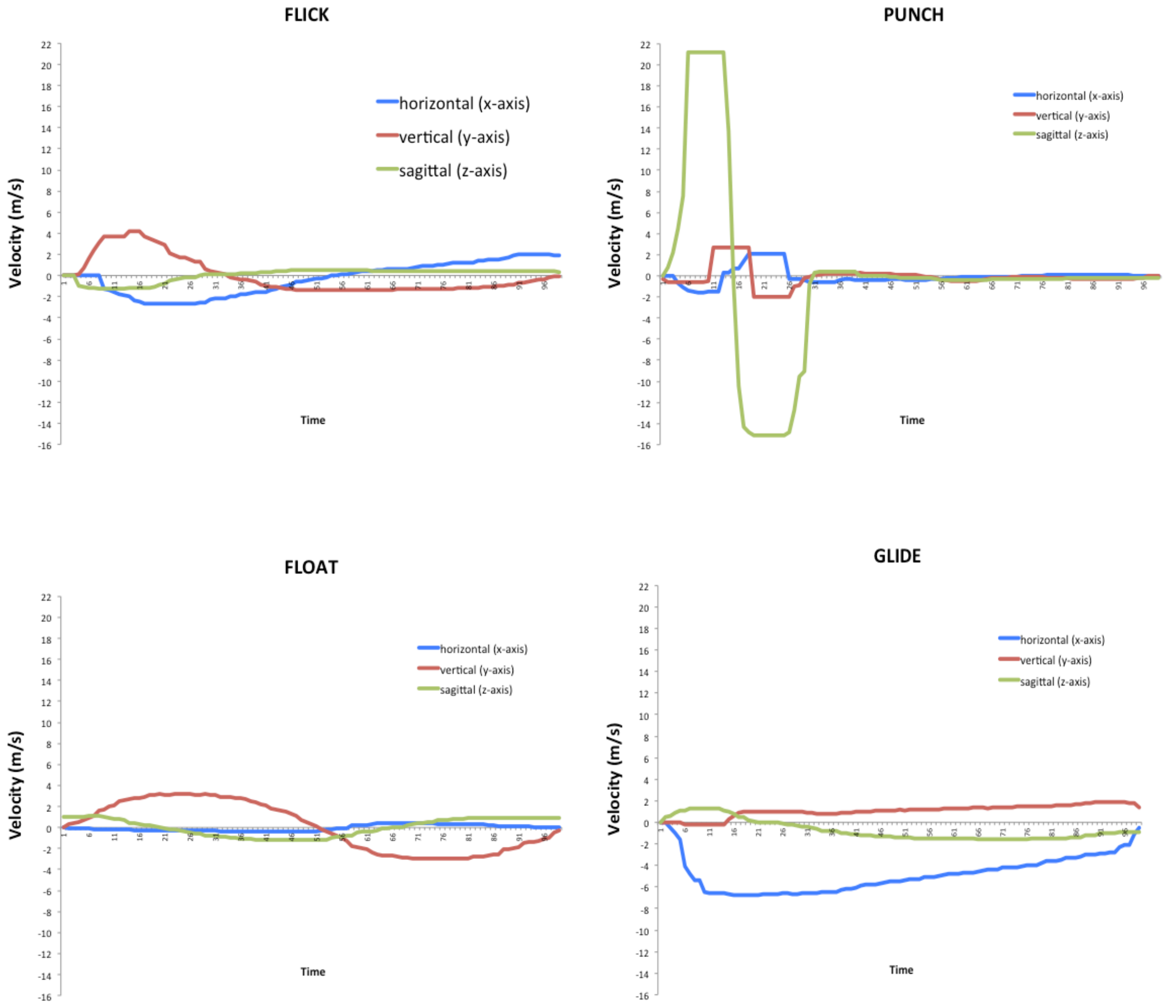


Figure 5. Horizontal (x-axis), vertical (y-axis) and sagittal (z-axis) velocity patterns for flick, punch, float and glide. The graphs represent velocity (m/s) on the y-axis, and duration on the x-axis (1000ms) for each of the four gestures.

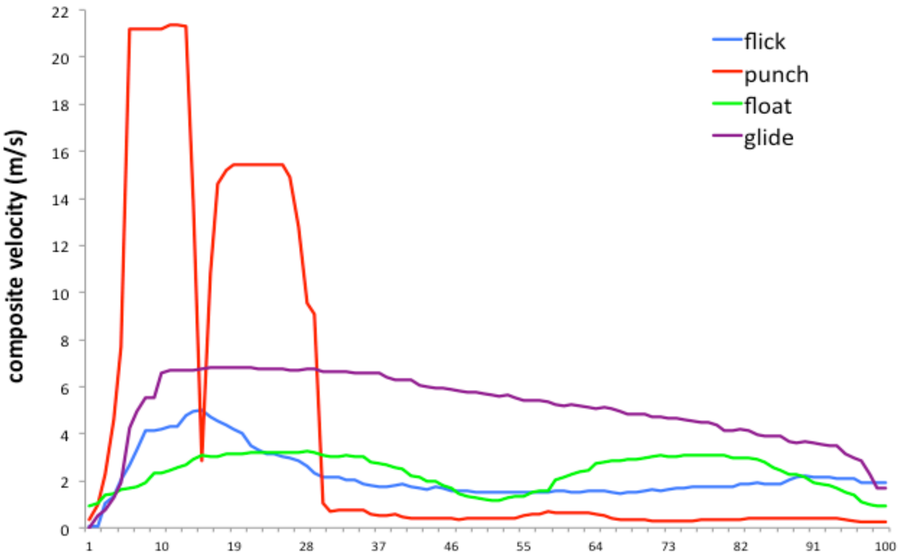


Figure 6. Composite velocity patterns superimposed for flick, punch, float, and glide. The graph represents velocity in m/s on the y-axis, and duration on the x-axis (1000ms).

	flick	punch	float	glide
Duration (ms)	540 ms	300 ms	1000 ms*	1000 ms*
Horizontal Displacement (HD)	*113.84 mm**	136.09 mm	21.82 mm	*570.15 mm**
Vertical Displacement (VD)	*105.41 mm**	137.97 mm	*154.41 mm**	137.50 mm
Sagittal Displacement (SD)	38.54 mm	*555.01 mm**	58.84 mm	127.90 mm
Composite Displacement (CD)	88.37 mm	328.82 mm	142.79 mm	143.57 mm
Mean Horizontal Velocity (HV)	0.15 m/s	0.12 m/s	0.03 m/s	*0.48 m/s**
Mean Vertical Velocity (VV)	*0.17 m/s**	0.16 m/s	*0.20 m/s**	0.11 m/s
Mean Sagittal Velocity (SV)	0.06 m/s	*1.30 m/s**	0.08 m/s	0.11 m/s
Mean Composite Velocity (MCV)	0.26 m/s	*1.33 m/s**	0.23 m/s	*0.51 m/s**
Composite Velocity Variation (CVV)	0.51 m/s	*0.66 m/s**	*0.63 m/s**	0.45 m/s
Mean Initial Velocity (MIV)	0.23 m/s	*1.22 m/s**	0.16 m/s	*0.31 m/s**
Maximum Elevation (ME)	*1550 mm**	1390 mm	*1523 mm**	1411 mm

Table 4. Summary of the calculated mathematical values for each of the movement features extracted from each gesture. For displacement and mean velocity the values marked with an asterisk represent the dominant axis with the bigger change. For all other motional parameters, the values marked with an asterisk represent the gestures with bigger change.

Table 4 shows that:

- In terms of *duration*, flick is longer than punch; and float and glide are longer than both flick and punch gestures.
- In terms of *displacement* and *mean velocity*, the "vertical axis" is the dominant axis for "flick", the "sagittal axis" is the dominant axis for "punch", the "vertical axis" is the dominant axis for "float", and the" horizontal axis" is the dominant axis for "glide".
- In terms of *initial and mean veloci*ty, punch is faster than flick; and glide is faster than float.
- Punch has increased *velocity variation* than flick; but float (with less initial/mean velocity) has increased *velocity variation* than glide.
- Flick has a bigger *maximum elevation* than punch; and float has a bigger *maximum elevation* than glide.

Analyses of vocal responses

Ratings by expert judges

In order to test whether participants varied their utterances for each gesture being observed, two independent judges familiar with the gestures, listened to and scored the audio-recordings of the vocal responses without any knowledge of the visual gestures, and predicted which of the four gestures gave rise to the produced /da/ sounds. We define accuracy as correct categorization of the sound samples into one of four movement categories, which are computed as percentages. Binomial tests were used to detect whether the percentages were above chance level, and Mann-Whitney tests were used to test whether the categorization accuracies for each experimental groups (experiment 2, 3, 4, & 5) were different from the control group (experiment 1), whenever appropriate.

For experiment 6 we adopted a different strategy for the perceptual task. Each judge listened to randomly ordered pre-test and post-test /da/ sounds for each of the four gestures, and guessed which of the two renditions were from post-test, ideally pointing towards a better representation of their respective gestures. Binomial tests were used to detect whether the percentages were above chance level.

Acoustic analyses of vocal responses

In order to further test whether the participants systematically varied the quality of their utterances, and whether/how the acoustic features corresponded to the observed motion features, we performed a detailed acoustic analysis of the sound data. Only the data from the second set of block was taken into consideration. The gestures were presented in two chunks of 5 repetitions, and the middle 3 /da/ sounds were used for the analyses, making it a total of 6 data points for each participant and each gesture. The acoustic analyses complement the perceptual task by expert judges.

Acoustic analyses of individual syllables were conducted using Praat speech processing software (Boersma and Weenink), as has been used by other researchers (Dalla Bella et al., 2007; Martinez-Castilla & Sotillo, 2008). Syllable and vowel boundaries were marked by hand by visually inspecting the waveform and spectrogram, and by listening to the segments.

The syllabic nuclei corresponding to the vowel /a/ were used for extracting fundamental frequency, pitch variability and formant frequencies. Syllable boundaries were used for extracting duration, amplitude and amplitude variability. Fundamental frequency (F0) of each vowel measured in Hz is computed using the Praat autocorrelation method. Vowel quality was specified as the difference between the second and first formant frequencies as done in Prieto and Ortega-Llebaria (2006).

The acoustic features that were extracted from each syllable (/da/) and what they represent are summarized in Table 5.

Acoustic Feature	**Label**	**Description**
duration (ms)	duration	The durational difference between syllable offset and onset in ms
mean amplitude (dB)	meanDB	Mean intensity measured in dB
amplitude variability (dB)	sdDB	Standard deviation of intensity in dB
fundamental frequency (Hz)	F0	Median perceived pitch in Hz
pitch variability (Hz)	sdPitch	Standard deviation of pitch in Hz
vowel quality (Hz)	F2-F1	The difference between the first and second formant frequencies in Hz

Table 5. The acoustic features extracted from the sound data, their label, what they represent, and how they were calculated. These were the six dependent measures in each of the six experiments.

The values of each acoustic parameter were, then, related to the motion characteristics of the gestures, as specified by the kinematic analysis of the gestures as shown above.

Specific hypotheses and statistical analyses methods

Hypothesized associations of acoustic and motional parameters:

Past studies applying a wide range of paradigms such as motion imagery (Eitan and Granot, 2006; Kohn and Eitan, 2009), drawings (Küssner and Leech-Wilkinson, 2014; Kussner et al., 2014), gestures (Nymoen et al., 2013) and forced choice discriminations (Walker, 1987) have shown strong associations between tempo and speed, loudness and muscular energy (especially along the z axis), and pitch and verticality. Based on the past literature summarized above, along with the movement data analyses we have presented, we have hypothesized the following cross-modal links and outcomes (Table 6), and tested them for significance.

Gesture	**Sound**
(1) higher duration	higher duration
(2) higher initial/mean velocity (energy)	higher amplitude
(3) higher velocity variation	higher amplitude variation
(4) higher vertical elevation	higher fundamental frequency
(5) higher vertical displacement/velocity	higher pitch variation
(6) higher mean velocity (energy)	higher F2-F1 difference

Table 6. Hypothesized cross-modal links

1) The most apparent of all relationships between motional and acoustic features is the association between tempo and the speed of human motion. In our study the visual gestures were presented at a fixed tempo (60 bpm), however the duration of the gestures varied. Therefore, we have predicted that participants would match the **duration** of the gesture with the **duration** of their speech sounds. Based on the kinematic analyses of duration, we predicted that abrupt gestures of flick and punch would elicit staccato responses, whereas the sustained gestures of float and glide would elicit legato responses (flick & punch vs. float & glide). Flick was expected to elicit longer responses than punch (flick vs. punch). Although both glide and float gestures take up the whole 100 ms to complete, the float gesture has a very slow velocity profile in which we had expected the participants to start their utterances later in time, which would result in a shorter float than glide (float vs. glide).
2) In the literature, muscular energy and forward-backward movement along the z-axis are linked to loudness. Muscular energy manifests itself as an increase in speed. So, we have predicted that **higher muscular energy / velocity** (especially along the z-

axis) would result in **higher mean amplitude** levels in the sound counterparts. Based on the kinematic analyses, we predicted that punch, with the highest mean and initial composite velocity (with the z-axis being the dominant axis) would lead to loudest response (punch vs. flick, punch vs. glide). Similarly we predicted that glide with higher mean and initial composite velocity would result in a louder response than float (glide vs. float). The velocity analyses were also in line with Laban's categorization of the gestures in terms of perceived weight. According to Laban categorization punch is produced with strongest weight/energy, followed by glide, and then followed by float and flick, both of which are of light weight/energy.

3) Based on the velocity (energy) and loudness dyad, we have hypothesized that **higher velocity variation** would be linked to **higher intensity variation**. Based on the kinematic analyses of composite and dominant axis velocity variation, we expected punch to arouse higher intensity variation than flick (punch vs. flick); and float to arouse higher intensity variation than glide (float vs. glide).
4) Past studies have shown that a strong association between pitch and height such that **higher elevation** in space (higher position on the vertical y-axis) is associated with **higher pitch** and vice versa. Based on the kinematic analyses, we have expected flick to result in higher mean fundamental frequency (F0) than punch (flick vs. punch), and we have expected float to result in higher mean F0 than glide (float vs. glide).
5) In the literature, spatial verticality (movement on y-axis) is linked with changes in pitch such that a rising and falling pitch contour results in a movement that is higher and lower in elevation. So, we have expected that the gestures with **higher vertical displacement and vertical velocity** to result in **higher overall pitch variability**. Based on the kinematic analyses of vertical displacement and velocity, we predicted

that flick would result in higher pitch variation than punch (flick vs. punch), and we predicted float would result in higher pitch variation than glide (float vs. glide).

6) Motion characteristics may also have an effect on the **vowel quality** as quantified by first (F1) and second formant frequency (F2) differences (as done in Prieto and Ortega-Llebaria, 2006). Formants are concentrations of acoustic energy around particular frequencies corresponding to a resonance in the vocal tract that gives the characteristics to different vowels. A bigger F2-F1 difference would indicate a vowel that is more to the front, and/or closed (high tongue), whereas a smaller F2-F1 difference would indicate a vowel that is more to the back and/or open (low longue). Previous literature did not specify any links between vowel quality and motional features, however Prieto and Ortega-Llebaria (2006) has observed that vocal "stress" (or pressure) changes the quality of the vowel [a] by decreasing the distance between F2 and F1. That is, the distance between F1 and F2 is smaller for stressed [a] than it is in its unstressed counterpart. Given the assumption that higher **overall energy/velocity** would lead to increased "stress" (pressure) on the /da/ sounds, we have predicted that punch (more stressed) would be produced with a smaller F2-F2 distance than flick (less stressed); and glide (more stressed) would be produced with a smaller F2-F2 distance than float (less stressed).

Lombard Effect states that increased loudness is also accompanied by an increase in pitch and pitch variability as well as a shift in formant center frequencies (Raphael et al., 2005). In other words, when people speak in noise, they not only increase their speech intensity, but there are also pitch related influences that accompany the increase in intensity. Past studies have reported an increase in voice F0 and F0 variability as a function of increased loudness, which involuntarily results from greater resistance by the vocal folds to increased airflow (Gramming et al. 1988; Raphael et al. 2007). Punch gesture was expected to arouse the most intense response, which could lead to increased pitch and pitch variation by itself. Therefore, due to the confounding Lombard effect, our expectations for the effects of motion verticality on F0 and pitch variation were less conclusive for the punch vs. flick pair.

Statistical Analyses methods

For experiment 1, A series of Linear mixed-effects models (Pinheiro & Bates, 2000) were used to examine the effect of viewing four different gestures on each of the dependent variable of acoustic parameters using the Proc Mixed procedure of SAS version 9.4 for Windows (SAS Institute, Cary, NC, USA). For each subject, random intercepts were added to allow accounting for the different baseline values for each subject. Mixed models also allowed for multiple observations from each individual to be taken into account rather than using average responses, which allowed more power. They also allowed for an unbalanced design where some of the utterances with pitch errors were excluded from the analysis. Six separate statistical models were constructed to examine each of the six dependent measures (i.e. duration, meanDB, sdDB, F0, sdPitch, F2-F1) across four viewing conditions (i.e. flick, punch, float, glide, fixed effect) with mean intensity values entered as covariate whenever necessary. In all models, mean intensity was taken into account as a covariate when the dependent variables were sdDB, F0, sdPitch and F2-F1. This was done because standard deviation of intensity is correlated by mean intensity; and pitch related measures of F0, sdPitch and F2-F1 are correlated by overall intensity as increased loudness is also accompanied by an increase in pitch and pitch variation (Raphael et al., 2005) due to the Lombard Effect. As the next step, planned comparisons were performed on specific gesture pairs (e.g. flick vs. punch; and float vs. glide) to explore whether the acoustic parameters are manipulated in a way predicted by the movement characteristics of the gestures.

For experiments 2,3,4 and 5 we compared the data from the experimental groups (e.g. weak instruction, musicians, PLD, and auditory feedback masking) with the data from the first study, which represents the original task performed by nonmusician college students (experiment 1). For these sets of analyses the data from the 1st experiment acted as a control, and we explored whether the weak instruction (experiment 2), having music background (experiment 3), watching point light displays (experiment 4), or absence of auditory feedback (experiment 5) affected the ability to map gesture on sound in similar or different ways.

Again, six separate statistical models were constructed to examine each of the six dependent measures (i.e. duration, meanDB, sdDB, F0, sdPitch, F2-F1) across four viewing conditions (i.e. flick, punch, float, glide, fixed effect) and across the two groups of interest (experimental vs. control). As the next step, planned comparisons were performed on specific gesture pairs (e.g. flick vs. punch; and float vs. glide) across the two groups. We have specifically looked for interaction effects of gesture type with group; where the experimental manipulation might have affected the way subjects vary their acoustic features.

For experiment 6 the linear-mixed effect model was constructed with a between-subjects fixed factor of group (i.e., motor practice group vs. visual practice group), a within-subjects fixed factor of time (pre-test scores vs. post test scores), and a within-subjects fixed factor of gesture type (flick, punch, float, glide). Specific pre-planned contrasts were estimated specifically to test the two pairs of interest (flick vs. punch) and (float vs. glide) as a function of time and group. Additionally, we have constructed specific two-way models for each of the four gestures in isolation, with time (pre-test vs. post-test) and group (motor vs. visual) in the model. We have specifically aimed for an interaction effect between time and group, which would entail a more pronounced difference in the acoustic measures at the post-test compared to pre-test (hence enhanced visual-to-auditory coupling) for one of the two practice groups. Then simple slopes were calculated for each of the practice group to evaluate the gain between pre-test and post-test scores.

CHAPTER III

RESULTS

Experiment 1. Cross-modal mapping of observed gestures onto vocal sounds

Ratings by expert judges

Categorization accuracies were computed for each of the 4 gestures by two independent judges familiar with the four Laban gestures. The inter-rater agreement for the two judges was ***κ***= 0.9, strength of which is considered “very good”. Cohen’s kappa is thought to be a more robust measure than simple percent agreement calculation, because it takes into account the agreement that might occur by chance (Cohen, 1960). The average accuracy scores were computed and the results are summarized in Table 7. Binomial tests were performed on the percentages of correct classifications (compared to chance level of %25), and the significant levels at the .05 level were marked with an asterisk. Figure 7 shows the distribution of the responses for each of the four gestures.

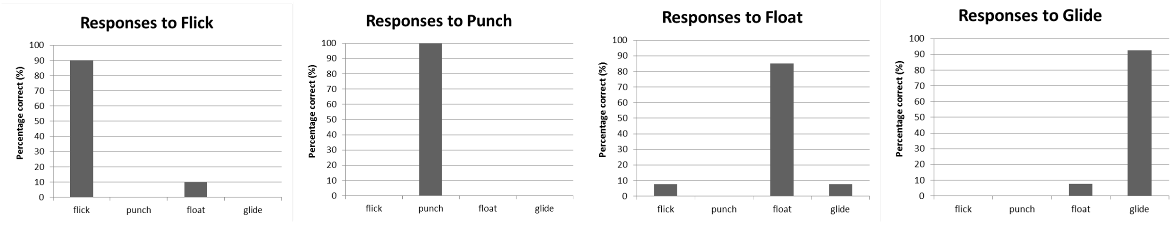


Figure 7. Distribution of categorization accuracies for experiment 1 (nonmusicians).

Categorization Accuracies	flick	punch	float	glide
Experiment 1 – nonmusicians	%90*	%100*	%85*	%92.5*

Table 7. Categorization accuracies for experiment 1 (nonmusicians). Scores were compared to chance level of (0.25) with a binomial test, and the significant levels at the .05 level were marked with an asterisk.

Acoustic analyses of vocal responses

Six separate linear mixed-effects models were run for each of the six dependent measures (duration, mean intensity, intensity variation, fundamental frequency, pitch variation and vowel quality) using the Proc Mixed procedure. The models included a within-subjects fixed factor of gesture type (flick, punch, float, glide), and mean intensity was used as a covariate whenever appropriate. For each subject, random intercepts were added to allow accounting for the different baseline values for each subject. The statistical formulas modelling six dependent measures as a function of the four different types of gestures are as follows:

(1) duration ~ gesture + (1|subject) + ε
(2) meanDB ~ gesture + (1|subject) + ε
(3) sdDB ~ gesture + meanDB + (1|subject) + ε
(4) F0 ~ gesture + meanDB + (1|subject) + ε
(5) sdPitch ~ gesture + meanDB + (1|subject) + ε
(6) F2-F1 ~ gesture + meanDB + (1|subject) + ε

Based on the hypothesized movement-sound associations as described in section *Associations of acoustic and motional parameters*, specific planned comparisons were computed and estimated. The Figure 8.a-e displays the output from the main models for each of the six dependent measures. The *p* values from the planned comparisons are displayed in a table below the figure (Table 8). Our approach was to specify hypothesis driven planned comparisons from the models, so no main effects are reported.

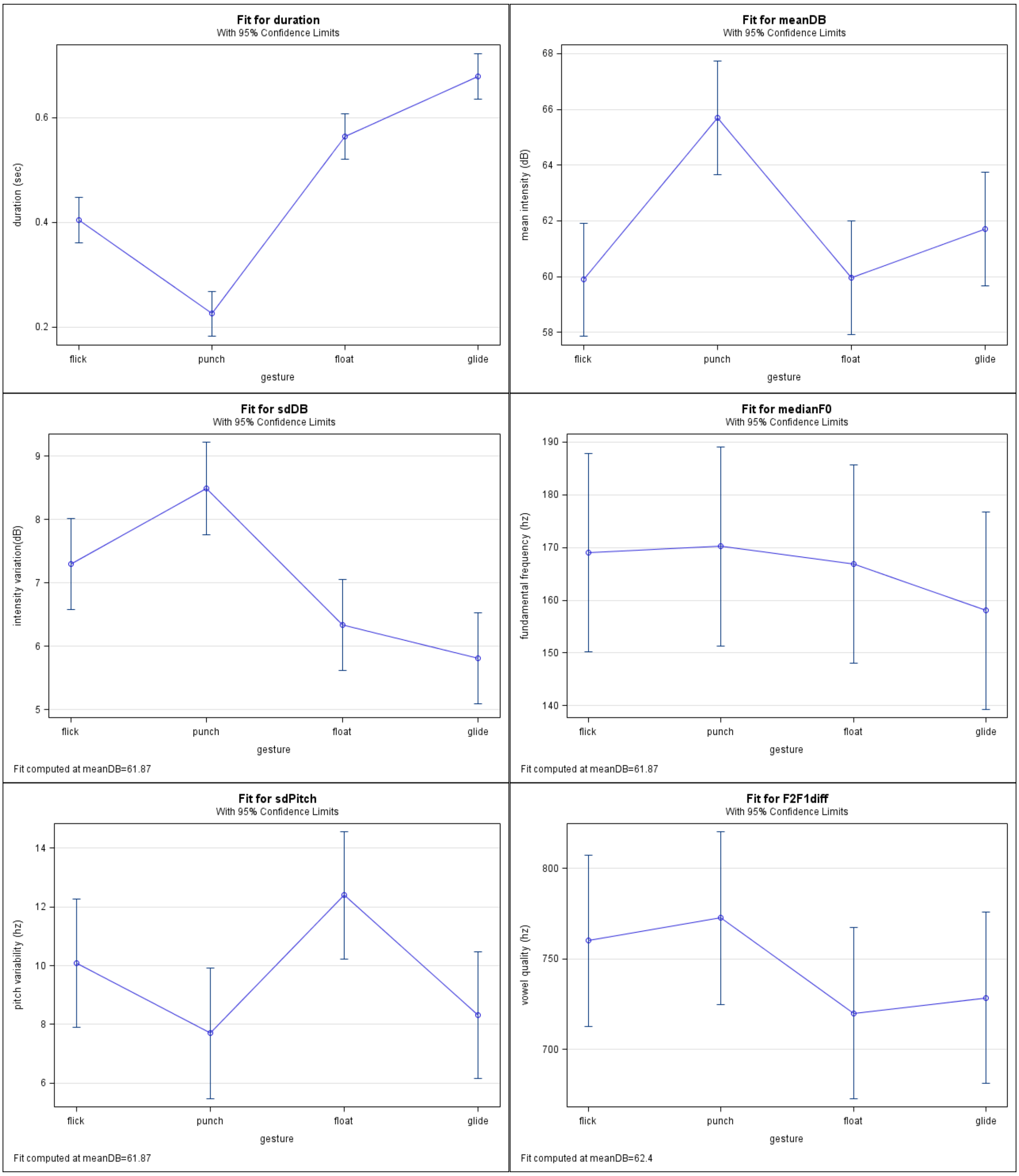


Figure 8a-f. Model outputs representing six acoustic features as a function of gesture type for experiment 1. Error bars represent confidence limits. These are computed as the normal (Wald) confidence limits for the linear predictor.

Experiment 1	**duration**	**mean intensity**	**intensity variation**	**fundamental frequency**	**pitch variation**	**vowel quality (F2-F1)**
flick vs. punch	β=.18, SE=.01, p=<.0001*	β=-5.81, SE=.37, p=<.0001*	β=-1.19, SE=0.23, p=<.0001*	n.s.	β=2.39, SE=.80, p=.0032*	n.s.
float vs. glide	β=-.16, SE=.01, p=<.0001*	β=-1.74, SE=.37, p=<.0001*	β=.52, SE=.19, p=.0072*	β=8.79, SE=1.68, p=<.0001*	β=4.07, SE=.68, p=<.0001*	n.s.
flick vs. float	β=-.16, SE=.01, p=<.0001*					
flick vs. glide	β=-.28, SE=.01, p=<.0001*					
punch vs. float		β=5.74, SE=.37, p=<.0001*				
punch vs. glide		β=3.99, SE=.37, p=<.0001*				

Table 8. The results from the planned comparisons for experiment 1. All predictions held except for the fundamental frequency for flick-punch pair, and the vowel quality for the flick-punch and float-glide pairs. Blue columns represent statistically significant pairs.

The participants in this experiment were nonmusician college students with no significant musical background; however they reliably varied the way they produced /da/ sounds by modifying their acoustic parameters in a way predicted from the movement analyses. Specifically they produced a punch /da/ shorter than flick, a flick /da/ shorter than float, and a float /da/ shorter than glide. Their punch /da/ was the loudest, and their glide /da/ was louder than their float. Their flick /da/ was less variable in intensity, higher in F0, and more variable in pitch compared to punch /da/. Similarly their float /da/ was less variable in intensity, higher in F0, and more variable in pitch compared to glide /da/. The expectation with regard to vowel quality did not hold.

Experiment 2. Effect of instruction on gesture sound coupling

The purpose of experiment 2 was to explore whether gesture sound coupling is based on an automatic or deliberate/strategic mapping of visual features onto acoustic features, by minimizing the instruction given to the participants so that the instruction did not involve the key word “match”.

Ratings by expert judges

The inter-rater agreement for the two judges was ***κ***= 0.48, strength of which is considered "moderate". This resulted mainly due to the overall lower accuracy scores, where for a considerable number of the participants the judges were randomly guessing. The categorization scores from the perceptual task comparing the control group (experiment 1) and weak instruction group (experiment 2) are summarized in Table 8, and Figure 8 shows the distribution of the responses for each of the four gestures. Binomial tests were performed on the percentages of correct classifications (compared to chance level of %25), and the significant levels were marked with an asterisk. For the 'weak instruction' group the *p* values were .041, .004, .041, and .102 for flick, punch, float and glide respectively. Mann-Whitney tests were computed for each of the gesture to compare the accuracy scores between the two groups. The differences were significant for each of the gesture at the levels of *p*=.003, *p*=.001, *p*=.009 and *p*=.001 for flick, punch, float and glide respectively. This suggests that weak instruction negatively affected the performance; however the performance levels were still above chance levels for flick, punch and float.

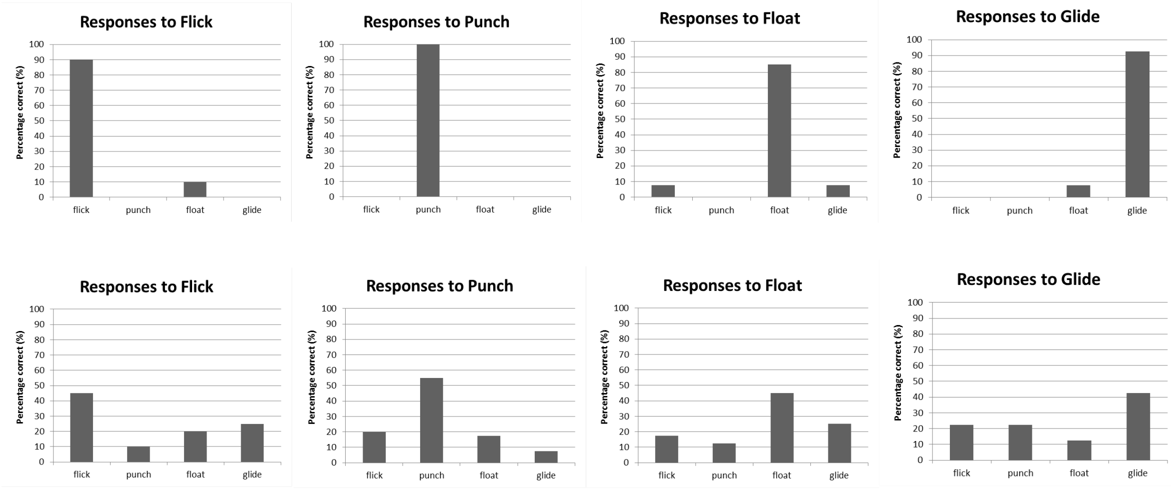


Figure 11. Distribution of categorization accuracies for experiment 1 (nonmusicians) in the upper panel, and experiment 2 (weak instruction) in the lower panel.

Categorization Accuracies	flick	punch	float	glide
Experiment 1 – nonmusicians	%90*	%100*	%85*	%92.5*
Experiment 2 –weak instruction	%45*	%55*	%45*	%42.5

Table 9. Categorization accuracies for experiment 1 (nonmusicians) on the upper panel, and experiment 2 (weak instruction) on the lower panel. Scores were compared to chance level of (0.25) with a binomial test, and the significant levels at the .05 level were marked with an asterisk.

Acoustic analyses of vocal responses

For experiments 2, 3, 4 and 5 we have compared the scores from the control group (experiment 1) with the scores from the experimental groups by adding a factor of group to the mixed model. The formulas modelling six dependent measures as a function of gesture and group are as follows:

(1) duration ~ gesture + group + (1|subject) + ε
(2) meanDB ~ gesture + group + (1|subject) + ε
(3) sdDB ~ gesture + group + meanDB + (1|subject) + ε
(4) F0 ~ gesture + group + meanDB + (1|subject) + ε
(5) sdPitch ~ gesture + group + meanDB + (1|subject) + ε
(6) F2-F1 ~ gesture + group + meanDB + (1|subject) + ε

Based on the hypothesized movement-sound associations, planned comparisons were computed and estimated within flick-punch and float-glide pairs. The Figure 10.a-e displays the output from the main models. The statistical results from the planned comparisons are displayed in a table below the figure (table 10).

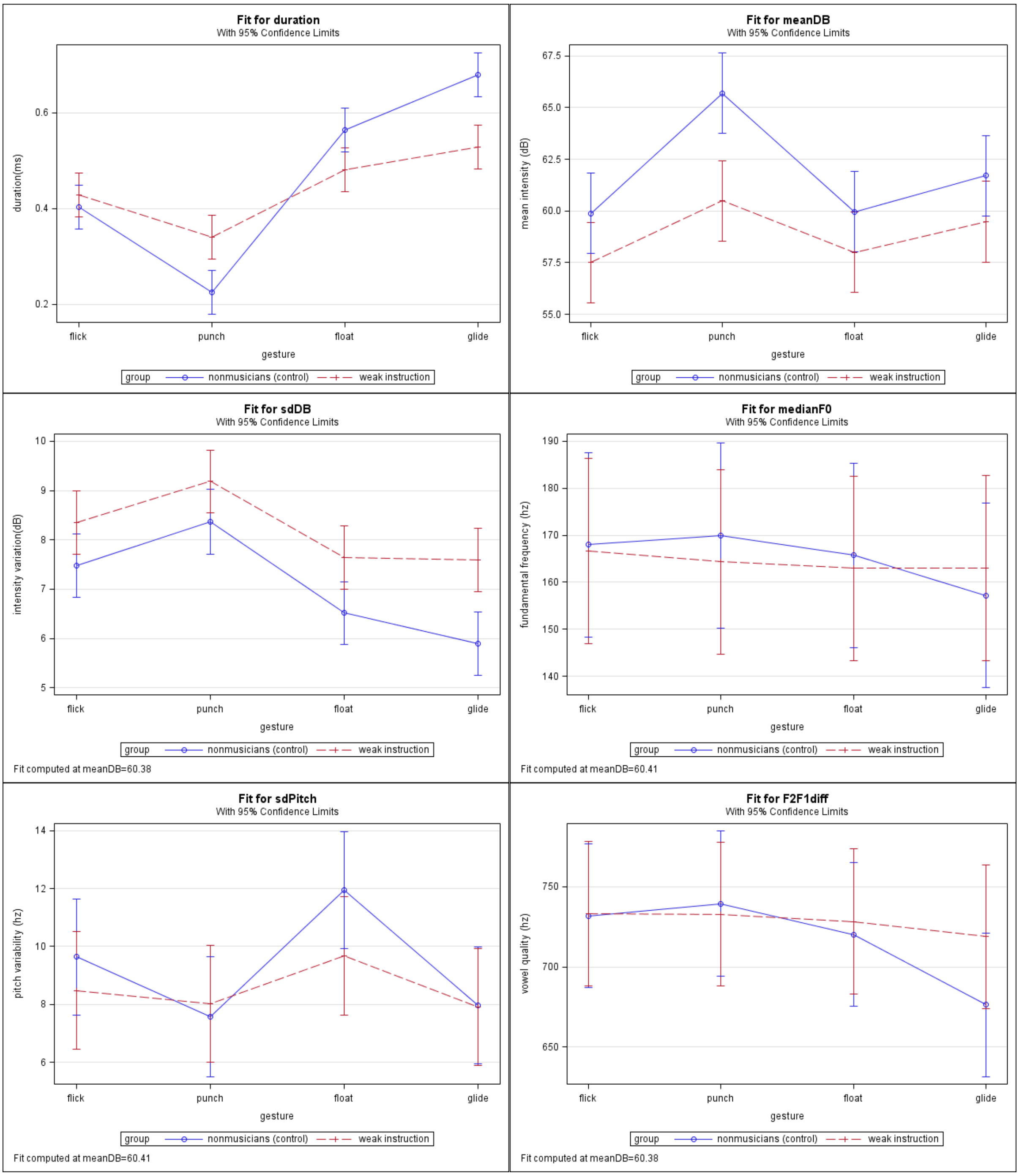


Figure 10a-f. Model outputs representing six acoustic features as a function of gesture type and group (control vs. weak instruction). Error bars represent confidence limits. These are computed as the normal (Wald) confidence limits for the linear predictor.

		duration	***mean intensity***	***intensity variation***	***fundamental frequency***	***pitch variation***	***vowel quality (F2-F1)***
flick vs. punch	*control*	β=.17, SE=.01, p=<.0001*	β=5.81, SE=.36, p=<.0001*	β=.89, SE=.22, p=<.0001*	*n.s.*	β=2.06, SE=.63, p=.0006*	*n.s.*
	weak instruction	β=.08, SE=.01, p=<.0001*	β=2.98, SE=.37, p=<.0001*	β=.83, SE=.21, p=<.0001*	β=2.3, SE=1.37, p=.0439*	*n.s.*	*n.s.*
	difference	β=.09, SE=.01, p=<.0001*	β=2.82, SE=.52, p=<.0001*	*n.s.*	β=-4.3, SE=1.89, p=.0232*	β=1.6, SE=.82, p=.0259*	*n.s.*
float vs. glide	*control*	β=.11, SE=.01, p=<.0001*	β=1.74, SE=.37, p=<.0001*	β=.61, SE=.20, p=.0014*	β=8.55, SE=1.32, p=<.0001*	β=3.98, SE=.57, p=<.0001*	β=43.8, SE=7.22, p=<.0001*
	weak instruction	β=-.04, SE=.01, p=.0005*	β=1.47, SE=0.38, p=<.0001*	*n.s.*	*n.s.*	β=1.76, SE=.59, p=.0015*	*n.s.*
	difference	β=.06, SE=.01, p=.0005*	*n.s.*	β=.60, SE=.29, p=.0267*	β=8.61, SE=1.88, p=<.0001*	β=2.21, SE=.82, p=.007*	β=34.2, SE=10.2, p=.0009*

Table 10. The results from the planned comparisons for experiment 2. Purple columns represent the acoustic variations that were achieved with weak instruction. Red columns represent the acoustic variations that were diminished by weak instruction compared to full instruction.

Results showed that with the weak instruction the participants still varied the basic parameters duration and amplitude across all four gestures, and the parameters of pitch were modified partially. Their 'punch' /da/ was overall shorter, louder, more variable in intensity, and lower in pitch than their 'flick' /da/, which was in line with our hypotheses of the cross-modal task. Similarly, their 'float' /da/ was overall softer and more variable in pitch then their 'glide' /da/, which was again, in line with our hypotheses. However, the differences were not as pronounced as in experiment 1 as shown by significant simple slope differences (representing an interaction effect). Therefore, weak instruction resulted in weaker modifications than the original instruction (experiment 1).

Experiment 3. Role of musical expertise in gesture sound coupling

Aim of experiment 3 was to explore whether musical expertise strengthens gesture sound mapping.

Ratings by expert judges

The inter-rater agreement for the two judges was ***κ***= 0.9, strength of which is considered "very good". The categorization scores from the perceptual task comparing the control group (experiment 1) and musicians group (experiment 3) are summarized in Table 11, and Figure 11 shows the distribution of the responses for each of the four gestures. Binomial tests were performed on the percentages of correct classifications (compared to chance level of %25), and results showed that all of the gestures could be predicted accurately by the judges (with a significance value of p < .0001). Mann-Whitney tests were computed for each of the gesture type to compare the accuracy scores between the two groups. The differences were not significant, suggesting the musicians and nonmusicians did not differ in their responses as evaluated by the perception of the judges.

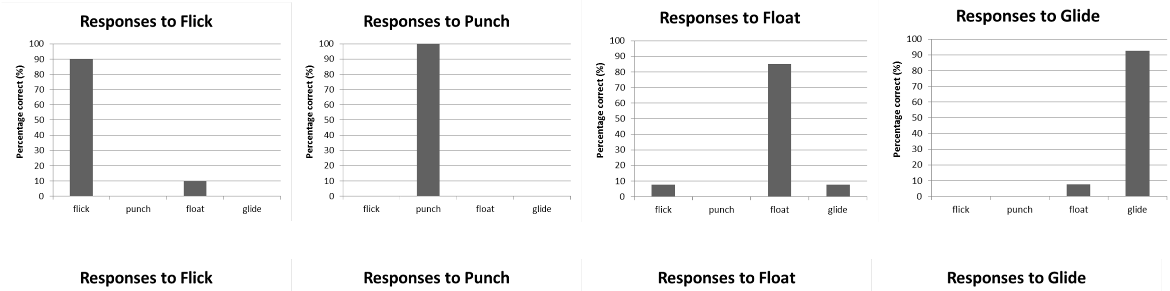


Figure 9. Distribution of categorization accuracies for experiment 1 (nonmusicians) on the upper panel, and experiment 3 (musicians) on the lower panel.

Categorization Accuracies	flick	punch	float	glide
Experiment 1 – nonmusicians	%90*	%100*	%85*	%92.5*
Experiment 3 –musicians	%92.5*	%100*	%85*	%77.5*

Table 11. Categorization accuracies for experiment 1 (nonmusicians) on the upper panel, and experiment 3 (musicians) on the lower panel. All categories were significant with a p value of <.0001 (marked with an asterisk). Mann-Whitney tests comparing the two groups were not significant at the p=.05 level for any of the gestures.

Acoustic analyses of vocal responses

The Figure 12.a-e displays the output from the main models. The statistical results from the planned comparisons are displayed in a table below the figure (table 12).

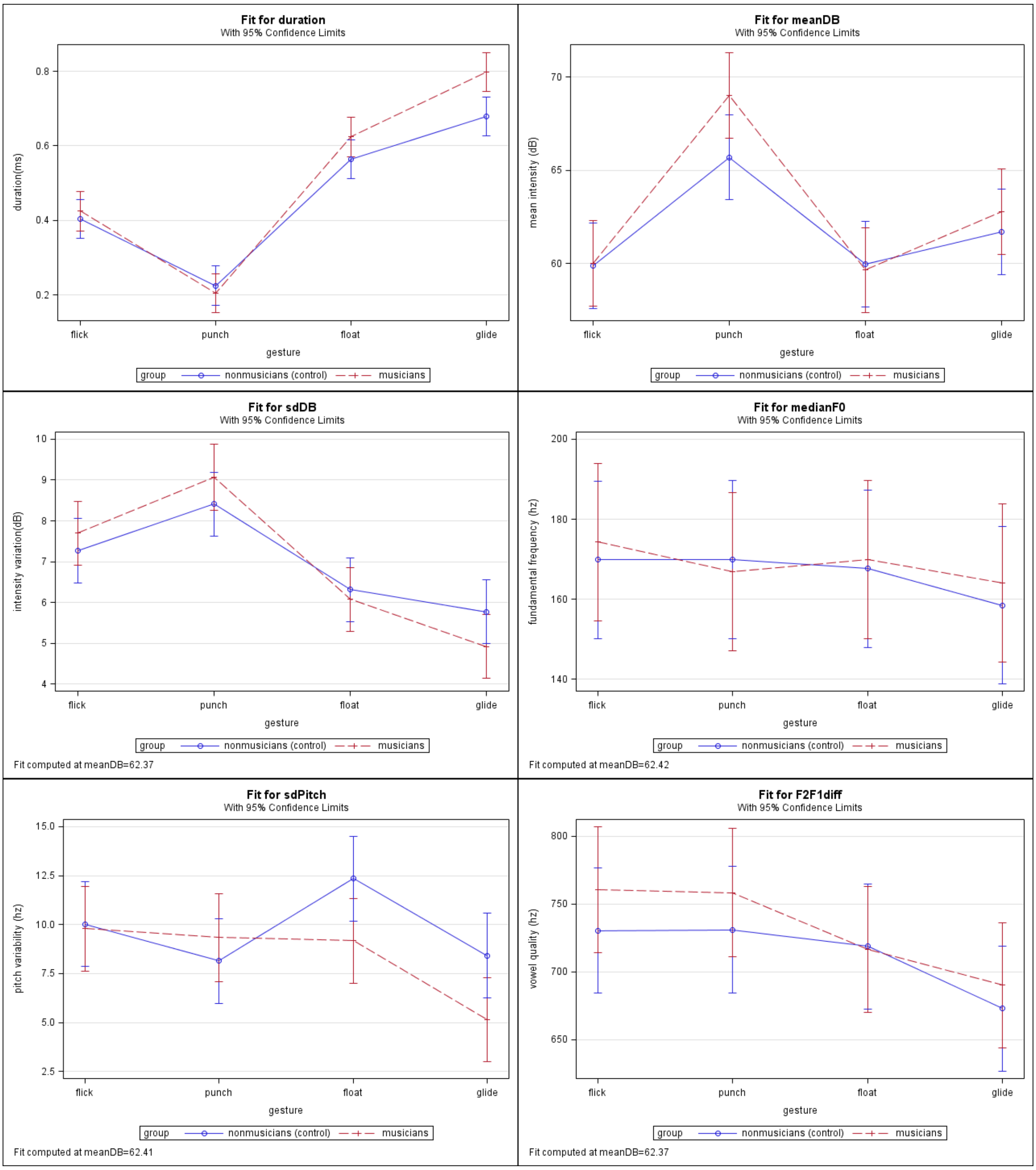


Figure 12a-f. Model outputs representing six acoustic features as a function of gesture type and group (control vs. musicians). Error bars represent confidence limits. These are computed as the normal (Wald) confidence limits for the linear predictor.

		duration	***mean intensity***	***intensity variation***	***fundamental frequency***	***pitch variation***	***vowel quality (F2-F1)***
flick vs. punch	*control*	*β*=.17, *SE*=.01, *p*=<.0001*	*β*=5.81, *SE*=.41, p=<.0001*	*β*=1.14, *SE*=.22, *p*=<.0001*	*n.s.*	*β*=1.87, *SE*=.70, *p*=.0041*	*n.s.*
	musicians	*β*=.22, *SE*=.01, *p*=<.0001*	*β*=9.01, *SE*=.42, *p*=<.0001*	*β*=1.36, *SE*=.25, *p*=<.0001*	*β*=7.34, *SE*=1.92, *p*=<.0001*	*n.s.*	*n.s.*
	difference	*n.s.*	*β*=3.20, *SE*=.59, p=<.0001*	*n.s.*	*β*=7.43, *SE*=2.24, *p*=.001*	*n.s.*	*n.s.*
float vs. glide	*control*	*β*=.11, *SE*=.01, *p*=<.0001*	*β*=1.74, *SE*=.42, *p*=<.0001*	*β*=.54, *SE*=.21, *p*=.0052*	*β*=9.11, *SE*=1.57, *p*=<.0001*	*β*=3.92, *SE*=.65, *p*=<.0001*	*β*=45.8, *SE*=8.61, *p*=<.0001*
	musicians	*β*=.17, *SE*=.01, *p*=<.0001*	*β*=3.14, *SE*=.42, *p*=<.0001*	*β*=1.15, *SE*=.21, *p*=<.0001*	*β*=5.88, *SE*=1.62, *p*=.0002*	*β*=4.01, *SE*=.67, *p*=<.0001*	*β*=26.4, *SE*=8.75, *p*=.0013*
	difference	*β*=.10, *SE*=.02, *p*=.0109*	*β*=1.40, *SE*=.59, *p*=.0186*	*β*=.60, *SE*=.29, *p*=.0407*	*n.s.*	*n.s.*	*n.s.*

Table 12. The results from the planned comparisons for experiment 3. Purple columns represent the acoustic variations that were achieved by musicians. Blue columns represent the acoustic variations that were enhanced with musical background (musicians).

As shown by significant simple slopes differences, musicians produced more pronounced differences in duration for float-glide pair; more pronounced differences in mean intensity for flick-punch and float-glide pairs, more pronounced differences in intensity variation for float-glide pair; and more pronounced differences in fundamental frequency for flick-punch pair. Moreover, musicians displayed fundamental frequency differences between flick and punch gestures, which were not previously observed within the control group. On the other hand, they did not show the pitch variation differences between flick and punch, which was shown with nonmusicians; however this could be due to the confounding effect of punch being produced at a higher intensity.

Experiment 4. Use of point light displays in gesture sound coupling

The purpose of experiment 4 was to explore whether gesture-sound links still remained when participants watched point light displays of the gestures to eliminate the featural information from the full body representations, and they observed the spatio-kinematic information only.

Ratings by expert judges

The inter-rater agreement for the two judges was **κ**= 0.9, strength of which is considered “very good”. The categorization scores from the perceptual task comparing the control group (experiment 1) and point light display group (experiment 4) are summarized in Table 10, and Figure 10 shows the distribution of the responses for each of the four gestures. Binomial tests were performed on the percentages of correct classifications (compared to chance level of %25), and the results showed that all categories were significant at the p < .000`1 level (marked with an asterisk). Mann-Whitney tests were computed for each of the gesture to compare the accuracy scores between the two groups. The differences were not significant, suggesting watching point light display representations did not result in deterioration as measured by perception of the judges.

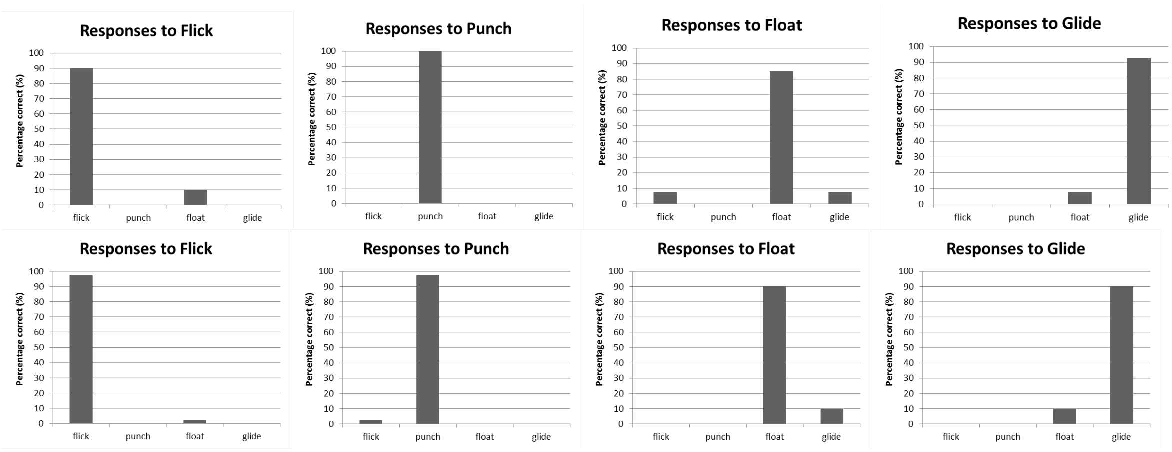


Figure 13. Distribution of categorization accuracies for experiment 1 (nonmusicians) on the upper panel, and experiment 4 (point light display) on the lower panel.

Categorization Accuracies	flick	punch	float	glide
Experiment 1 – nonmusicians	%90*	%100*	%85*	%92.5*
Experiment 4 –point light display	%97.5*	%97.5*	%90*	%90*

Table 13. Categorization accuracies for experiment 1 (nonmusicians) on the upper panel, and experiment 4 (point light display) on the lower panel. All categories were significant with a p value of <.0001 (marked with an asterisk). Mann-Whitney tests comparing the two groups were not significant at the p=.05 level for any of the gestures.

Acoustic analyses of vocal responses

The Figure 14.a-f displays the output from the main models. The statistical results from the planned comparisons are displayed in a table below the figure (table 14).

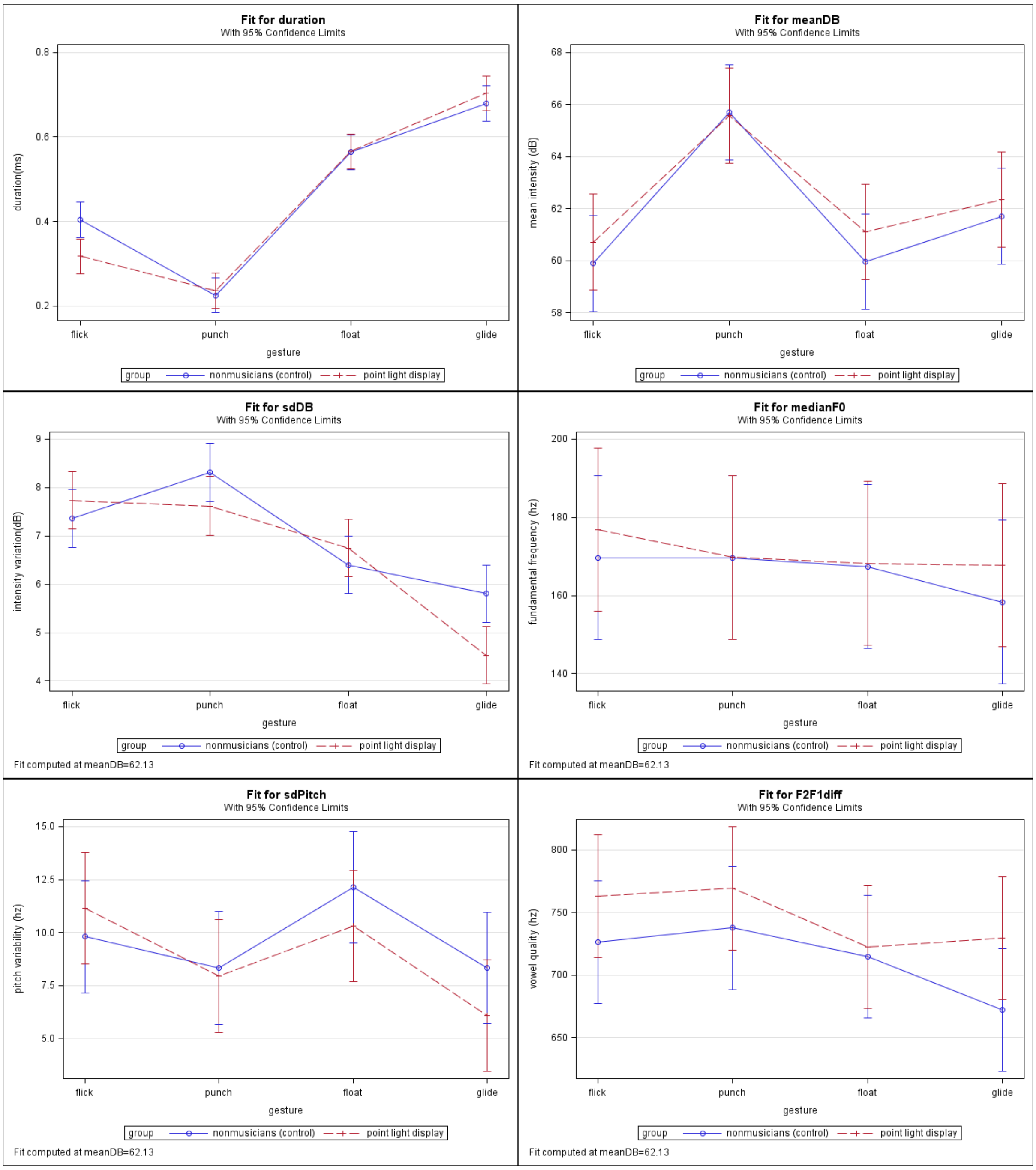


Figure 14a-f. Model outputs representing six acoustic features as a function of gesture type and group (control vs. point light display). Error bars represent confidence limits. These are computed as the normal (Wald) confidence limits for the linear predictor.

		duration	mean intensity	intensity variation	fundamental frequency	pitch variation	vowel quality (F2-F1)
flick vs. punch	control	β=.17, SE=.01, p=<.0001*	β=5.81, SE=.34, p=<.0001*	β=.95, SE=.22, p=<.0001*	n.s.	β=1.47, SE=.77, p=.0288*	n.s.
	PLDs	β=.08, SE=.01, p=<.0001*	β=4.85, SE=.34, p=<.0001*	n.s.	β=7.10, SE=1.69, p=<.0001*	β=3.21, SE=.74, p=<.0001*	n.s.
	difference	β=.09, SE=.02, p=<.0001*	n.s.	β=1.06, SE=.27, p=.0001*	β=-7.1, SE=2.17, p=.0011*	n.s.	n.s.
float vs. glide	control	β=.11, SE=.01, p=<.0001*	β=1.74, SE=.34, p=<.0001*	β=.59, SE=.19, p=.0014*	β=9.13, SE=1.56, p=<.0001*	β=3.80, SE=.69, p=<.0001*	β=42.6, SE=7.78, p=<.0001*
	PLDs	β=.13, SE=.01, p=<.0001*	β=1.23, SE=.34, p=.0002*	β=2.22, SE=.19, p=<.0001*	n.s.	β=4.23, SE=.68, p=<.0001*	n.s.
	difference	n.s.	n.s.	β=-1.62, SE=.27, p=<.0001*	β=8.68, SE=2.18, p=<.0001*	n.s.	β=49.7, SE=10.8, p=<.0001*

Table 14. The results from the planned comparisons for experiment 4. Purple columns represent the acoustic variations that were achieved by PLD group. Blue columns represent the acoustic variations that were enhanced with the point light stimuli. Red columns represent the acoustic variations that were diminished with the point light stimuli.

Simple slope analyses showed that the participants watching point light displays produced reliable variations in their vocalizations across the predicted pairs of gestures. As different from the control group, they produced reliable F0 differences between flick and punch pair (which was hypothesized in the beginning of the study), however they did not show differences in F0 and vowel quality between float and glide, and they did not show differences in intensity variation between flick and punch. As shown by significant simple slopes differences, they showed slightly enhanced difference effects between float and glide for intensity variation. They also showed slightly diminished duration differences between flick and punch, but this effect turned out to be a spurious effect caused by the absence of finger markers during the recording session of flick gesture in motion capture.

Experiment 5. Role of self-produced auditory feedback in gesture sound coupling

The purpose of experiment 5 was to explore the role played by auditory feedback in the cross-modal mapping of gesture and sound by masking the auditory feedback available to the participants.

Ratings by expert judges

The inter-rater agreement for the two judges was **κ**= 0.82, strength of which is considered “very good”. The categorization scores from the perceptual task comparing the control group (experiment 1) and auditory masking group (experiment 5) are summarized in Table 11, and Figure 11 shows the distribution of the responses for each of the four gestures. Binomial tests were performed on the percentages of correct classifications (compared to chance level of %25), and the results showed that all categories were significant at the *p* < .000`1 level (marked with an asterisk). Mann-Whitney tests were computed for each of the gesture to compare the accuracy scores between the two groups. The differences were not significant; suggesting being deprived of auditory feedback did not result in deterioration as measured by perception of the judges.

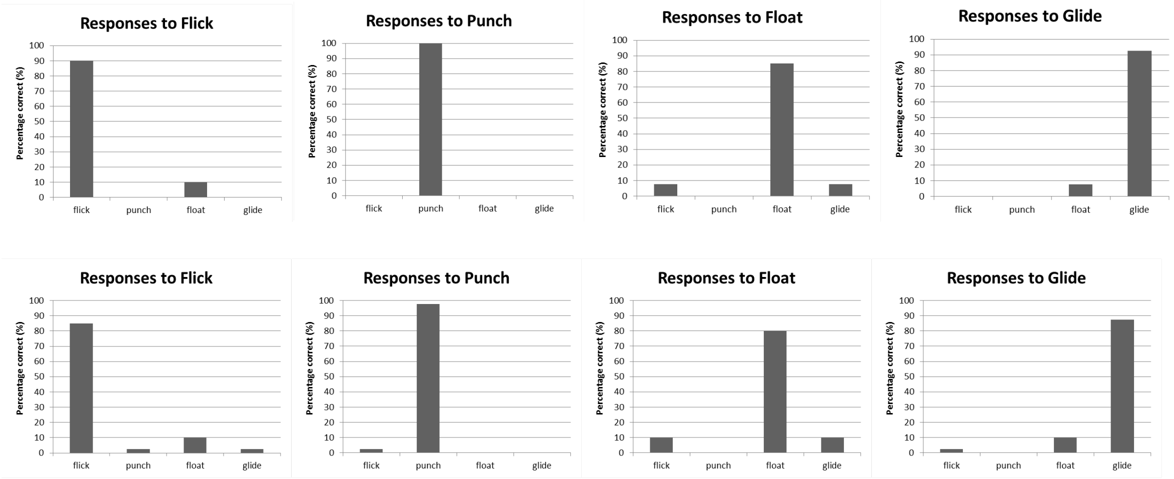


Figure 15. Distribution of categorization accuracies for experiment 1 (nonmusicians) on the upper panel, and experiment 5 (auditory masking) on the lower panel.

Categorization Accuracies	flick	punch	float	glide
Experiment 1 – nonmusicians	%90*	%100*	%85*	%92.5*
Experiment 5 –auditory masking	%85*	%97.5*	%80*	%87.5*

Table 15. Categorization accuracies for experiment 1 (nonmusicians) on the upper panel, and experiment 5 (auditory masking) on the lower panel. All categories were significant with a *p* value of <.0001 (marked with an asterisk). Mann-Whitney tests comparing the two groups were not significant at the *p*=.05 level for any of the gestures.

Acoustic analyses of vocal responses

The Figure 16.a-e displays the output from the main models. The statistical results from the planned comparisons are displayed in a table below the figure (table 16).

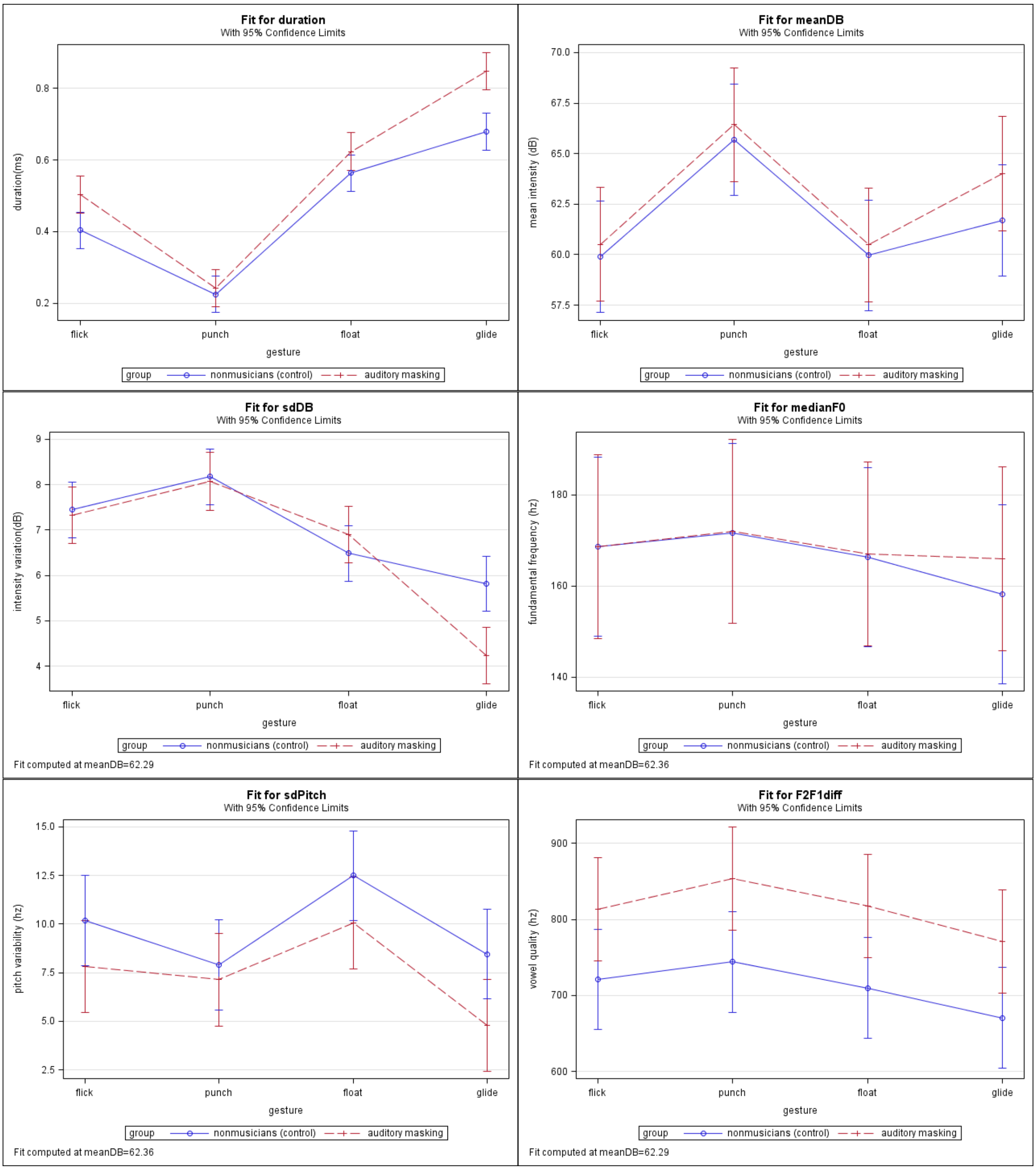


Figure 16a-e. Model outputs representing six acoustic features as a function of gesture type and group (control vs. masked auditory feedback). Error bars represent confidence limits. These are computed as the normal (Wald) confidence limits for the linear predictor.

		duration	***mean intensity***	***intensity variation***	***fundamental frequency***	***pitch variation***	***vowel quality (F2-F1)***
flick vs. punch	*control*	*β*=.17, *SE*=.01, *p*=<.0001*	*β*=5.81, *SE*=.39, p=<.0001*	*β*=.73, *SE*=.23, *p*=.0011*	*n.s.*	*β*=2.27, *SE*=.68, *p*=.0005*	*n.s.*
	Masked AF	*β*=.26, *SE*=.01, *p*=<.0001*	*β*=5.91, *SE*=.40, *p*=<.0001*	*β*=.74, *SE*=.24, *p*=.0012*	*n.s.*	*n.s.*	*n.s.*
	difference	*β*=-.08, *SE*=.02, *p*=.0011*	*n.s.*	*n.s.*	*n.s.*	*β*=1.60, *SE*=.89, *p*=.0746*	*n.s.*
float vs. glide	*control*	*β*=.11, *SE*=.01, *p*=<.0001*	*β*=1.74, *SE*=.39, *p*=<.0001*	*β*=.66, *SE*=.22, *p*=.0014*	*β*=8.25, *SE*=1.83, *p*=<.0001*	*β*=4.04, *SE*=.63, *p*=<.0001*	*β*=39.2, *SE*=7.87, *p*=<.0001*
	Masked AF	*β*=.22, *SE*=.01, *p*=<.0001*	*β*=3.52, *SE*=.41, *p*=.0002*	*β*=2.66, *SE*=.23, *p*=<.0001*	*n.s.*	*β*=5.27, *SE*=.67, *p*=<.0001*	*β*=46.0, *SE*=8.39, *p*=<.0001*
	difference	*β*=-.10, *SE*=.02, *p*=<.0001*	*β*=-1.78, *SE*=.57, *p*=.002*	*β*=-1.99, *SE*=.31, *p*=<.0001*	*β*=7.21, *SE*=2.63, *p*=.0063*	*n.s.*	*n.s.*

Table 16. The results from the planned comparisons for experiment 5. Purple columns represent the acoustic variations that were achieved with masked auditory feedback. Blue columns represent the acoustic variations that were enhanced with masked auditory feedback. Red columns represent the acoustic variation that was diminished with masked auditory feedback.

The participants in the auditory feedback-masking group reliably modified all of the acoustic features as hypothesized except for fundamental frequency for both pairs of gestures of interest, pitch variation and vowel quality for flick-punch pair only. Among these, only pitch variation difference was not achieved, as different from the control group. Duration and energy/velocity from the movements were not only reliably mapped onto duration and intensity in the accompanying sound counterparts, but the effects were even "enhanced" when auditory feedback was absent as shown by the significant simple slope differences. The effects were more pronounced for duration for the flick and punch pair; and they were more pronounced for duration, mean intensity and intensity variation for the float and glide pair.

Experiment 6. Effect of gestural motor practice in gesture sound coupling

The purpose of experiment 6 was to explore whether motor practice of observed gestures enhances the visual-to-auditory coupling.

Ratings by expert judges

Since the accuracy scores were at a ceiling level for the experiment 1, we adopted a forced choice paradigm for experiment 6 in the perceptual task. Each judge listened to randomly ordered pre-test and post-test performances for each of the four gesture, and guessed which of the two renditions were from post-test, representing better performance.

The inter-rater agreements for the two judges were computed for each gesture separately. They were κ= 0.74, 0.91, 0.66 and 0.66 for flick, punch, float and glide respectively for the motor practice group; and they were κ= 0.57, 0.74, 0.66 and 0.65 for the visual practice group. A score between 0.41 and 0.6 is considered "moderate"; a score between 0.61-0.8 is considered good; and a score between 0.81-1.00 is considered "very good" of strength. The percent scores representing better performance at the post-test (compared to pre-test) comparing the experimental group (motor practice) and the control group (visual practice) are summarized in Table 17, and Figure 17 shows the distribution of the responses for each of the four gestures. Binomial tests were performed on the percentages of correct identification of better post-test performance (compared to chance level of %50), and the significant levels were marked with an asterisk. The binomial tests were p=.01, p<.0001, p=.003, and p=.07 for flick, punch, float and glide respectively for the motor practice group; and they were p=.03, p=.27, p=.07, and p=.42, for the visual practice group.

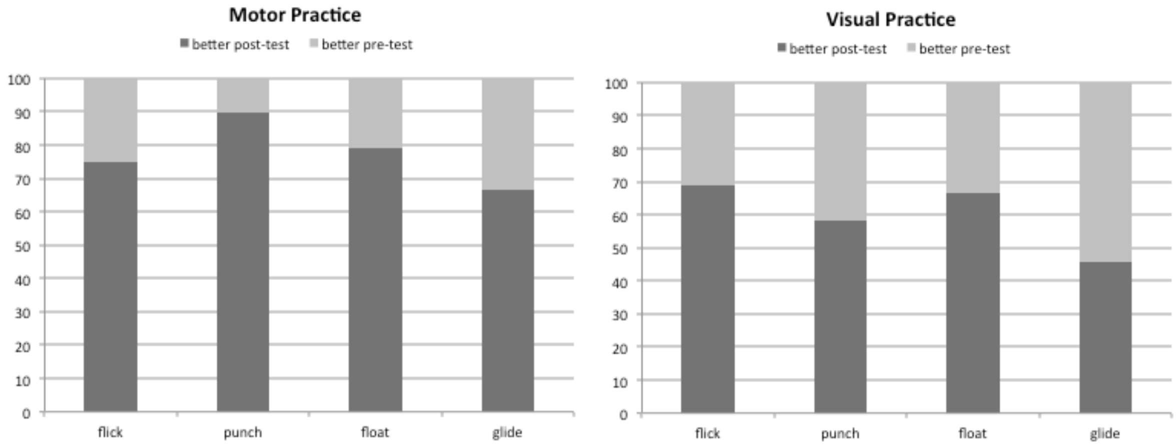


Figure 17. Distribution of percent accuracies for the motor practice group on the left and visual practice group on the right. Dark portion of the bars represents identification of post-test as better performance, and light portion represents identification of pre-test as better performance, as perceptually detected by the judges.

Percent Accuracies for Experiment 6	flick	punch	float	glide
Motor Practice Group	%75*	%90*	%80*	%66$^{p=.07}$
Visual Practice Group	%68*	%58	%66$^{p=.07}$	%45

Table 17. The percent scores representing better performance at the post-test (compared to pre-test) comparing the experimental group (motor practice) and the control group (visual practice) in Experiment 6. Significant levels are marked with an asterisk.

Acoustic analyses of vocal responses

Part 1

A series of Linear mixed-effects models were run for each of the six dependent measures (duration, mean intensity, intensity variation, fundamental frequency, pitch variation and vowel quality) using the Proc Mixed procedure of SAS version 9.4 for Windows (SAS Institute, Cary, NC, USA). The models included a between-subjects fixed factor of group (i.e., motor practice group vs. visual practice group), a within-subjects fixed factor of time (pre-test scores vs. post test scores), and a within-subjects fixed factor of gesture type (flick, punch, float, glide). For each subject, random intercepts were added to allow accounting for the different baseline values for each subject. The statistical formulas modelling six dependent measures as a function of gesture and time were as follows:

(1) duration ~ gesture + time + group + (1|subject) + ε
(2) meanDB ~ gesture + time + group + (1|subject) + ε
(3) sdDB ~ gesture + meanDB + time + group + (1|subject) + ε
(4) F0 ~ gesture + meanDB + time + group + (1|subject) + ε
(5) sdPitch ~ gesture + meanDB + time + group + (1|subject) + ε
(6) F2-F1 ~ gesture + meanDB + time + group + (1|subject) + ε

From the models, planned complex comparisons were estimated for two pairs of interest (flick vs. punch) and (float vs. glide). First step included comparing the two gestures of interest in each pair during pre-test and post-test. Second step included comparing the pre-test to post-test difference scores for motor practice and visual practice groups separately. And finally, the third step included comparing the gains from pre-test to post-test across motor practice and visual practice groups (representative of a three way interaction). The Figure 18 a-f displays the output from the models for each of the six dependent measures. The estimates, standard errors and *p* values from the planned complex comparisons are displayed in Table 18 a-f below each figure.

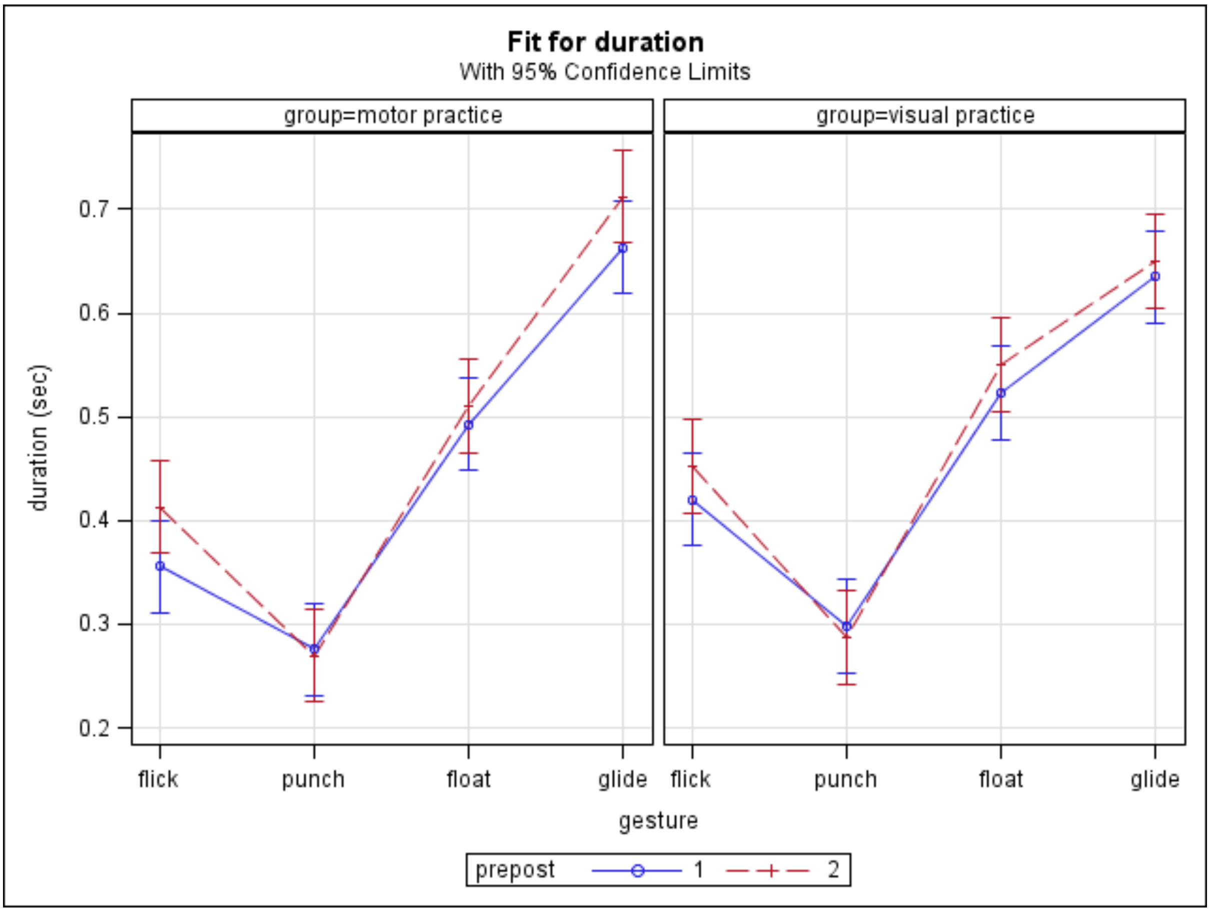


Figure 18a. The model output for "duration".

		Motor Practice		*Visual Practice*
Flick vs. Punch	*pre-test (time1)*	β=.07, *SE*=.01, *p*=<.0001*		β=.12, *SE*=.01, *p*=<.0001*
	post-test (time2)	β=.14, *SE*=.01, *p*=<.0001*		β=.16, *SE*=.01, *p*=<.0001*
	pre post difference	β=.06, *SE*=.01, *p*=.0004*		β=.04, *SE*=.01, *p*=.0129*
	*group*gesture*time*		*n.s.*	
Float vs. Glide	*pre-test (time1)*	β=.17, *SE*=.01, *p*=<.0001*		β=.11, *SE*=.01, *p*=<.0001*
	post-test (time2)	β=.20, *SE*=.01, *p*=<.0001*		β=.09, *SE*=.01, *p*=<.0001*
	pre post difference	β=.003, *SE*=.01, *p*=.0493*		*n.s.*
	*group*gesture*time*		β=.04, *SE*=.02, *p*=.0530*	

Table 18a. The estimates, standard errors and *p* values from the planned complex comparisons for 'duration'. Blue columns represent significant values. Of particular interest are the group*gesture*time interactions, which would indicate that the differences in duration between the two gestures of each pair were enhanced from pre to post-test at a greater degree for the motor practice group compared to visual practice group.

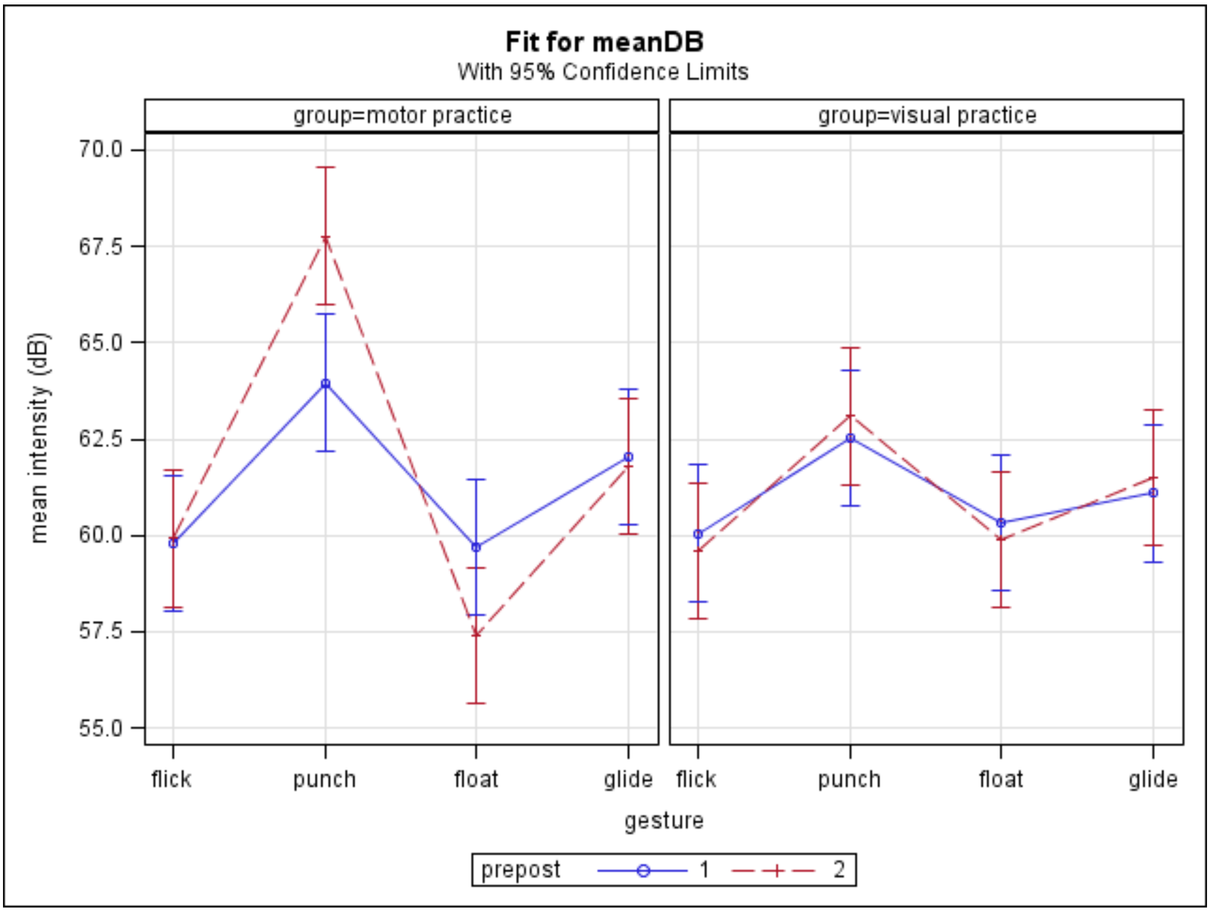


Figure 18b. The model output for “mean intensity”.

		Motor Practice		***Visual Practice***
Flick vs. Punch	*pre-test (time1)*	*β*=4.17, *SE*=.34, *p*=<.0001*		*β*=2.48, *SE*=.34, *p*=<.0001*
	post-test (time2)	*β*=7.85, *SE*=.34, *p*=<.0001*		*β*=3.50, *SE*=.34, *p*=<.0001*
	pre post difference	*β*=3.67, *SE*=.48, *p*=<.0001*		*β*=1.02 *SE*=.48, *p*=.0182*
	*group*gesture*time*		*β*=2.65, *SE*=.68, *p*=<.0001*	
Float vs. Glide	*pre-test (time1)*	*β*=2.31, *SE*=.34, *p*=<.0001*		*β*=.077, *SE*=.34, *p*=.011*
	post-test (time2)	*β*=4.39, *SE*=.341, *p*=<.0001*		*β*=1.61, *SE*=.34, *p*=<.0001*
	pre post difference	*β*=2.07, *SE*=.48, *p*=<.0001*		*β*=.83, *SE*=.48, *p*=.0418*
	*group*gesture*time*		*β*=1.24, *SE*=.68, *p*=.0350*	

Table 18b. The estimates, standard errors and *p* values from the planned complex comparisons for ‘mean intensity’. Blue columns represent significant values. Of particular interest are the group*gesture*time interactions, which would indicate that the differences in mean intensity between the two gestures of each pair are enhanced from pre to post-test at a greater degree for the motor practice group compared to visual practice group.

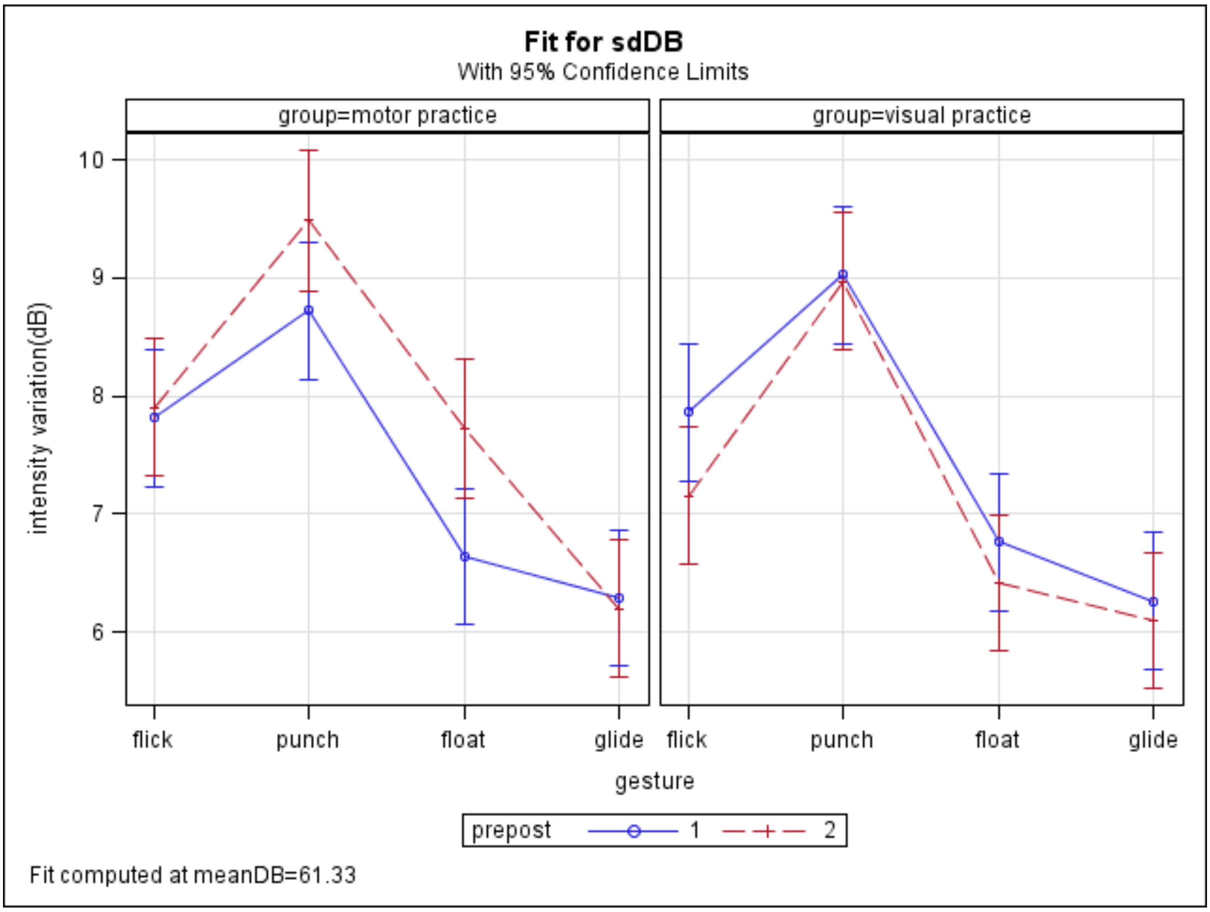


Figure 18c. The model output for "intensity variation".

		Motor Practice		*Visual Practice*
Flick vs. Punch	*pre-test (time1)*	β=.90, SE=.20, p=<.0001*		β=1.16, SE=.19, p=<.0001*
	post-test (time2)	β=1.58, SE=.21, p=<.0001*		β=1.81, SE=.20, p=<.0001*
	pre post difference	β=0.67, SE=.28, p=.0079*		β=.65 SE=.27, p=.0097*
	*group*gesture*time*		*n.s.*	
Float vs. Glide	*pre-test (time1)*	β=.34, SE=.19, p=.0388*		β=.49, SE=.19, p=.0055*
	post-test (time2)	β=1.51, SE=.20, p=<.0001*		β=.31, SE=.19, p=.0543*
	pre post difference	β=1.17, SE=.27, p=<.0001*		*n.s.*
	*group*gesture*time*		β=1.35, SE=.39, p=.0003*	

Table 18c. The estimates, standard errors and *p* values from the planned complex comparisons for 'intensity variation'. Blue columns represent significant values. Of particular interest are the group*gesture*time interactions, which would indicate that the differences in intensity variation between the two gestures of each pair are enhanced from pre to post-test at a greater degree for the motor practice group compared to visual practice group.

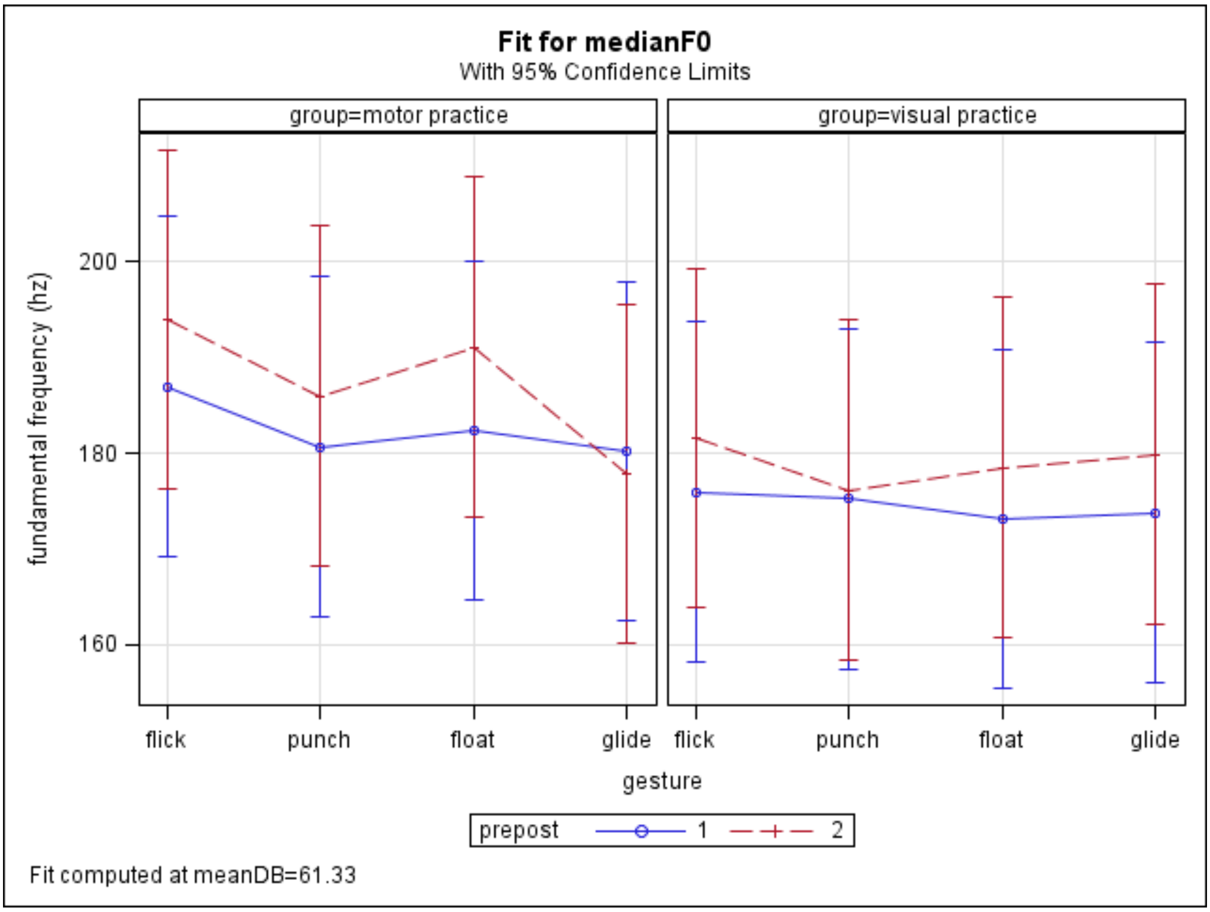


Figure 18d. The model output for "fundamental frequency".

		Motor Practice		*Visual Practice*
Flick vs. Punch	*pre-test (time1)*	*β*=6.26, *SE*=1.55, *p*=<.0001*		*n.s.*
	post-test (time2)	*β*=7.99, *SE*=1.67, *p*=<.0001*		*β*=5.44, *SE*=1.54, *p*=.0002*
	pre post difference	*n.s.*		*β*=4.69 *SE*=2.14, *p*=.0143*
	*group*gesture*time*		*n.s.*	
Float vs. Glide	*pre-test (time1)*	*n.s.*		*n.s.*
	post-test (time2)	*β*=13.2, *SE*=1.55, *p*=<.0001*		*n.s.*
	pre post difference	*β*=11.1, *SE*=2.13, *p*=<.0001*		*n.s.*
	*group*gesture*time*		*β*=11.83, *SE*=3, *p*=<.0001*	

Table 18d. The estimates, standard errors and *p* values from the planned complex comparisons for 'fundamental frequency'. Blue columns represent significant values. Of particular interest are the group*gesture*time interactions, which would indicate that the differences in fundamental frequency between the two gestures of each pair are enhanced from pre to post-test at a greater degree for the motor practice group compared to visual practice group.

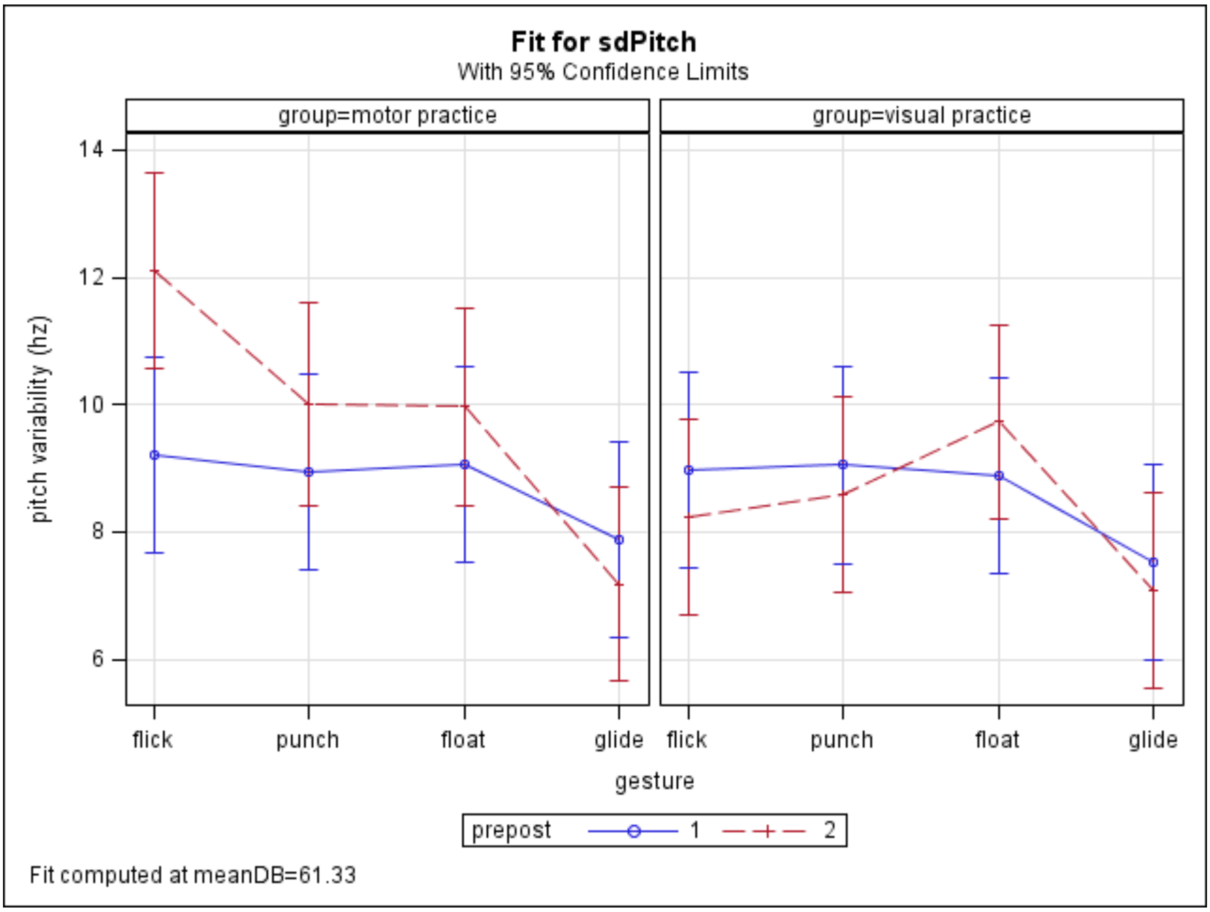


Figure 18e. The model output for “pitch variability”.

		Motor Practice		*Visual Practice*
Flick vs. Punch	*pre-test (time1)*	n.s.		n.s.
	post-test (time2)	β=2.1, SE=.62, p=.0004*		n.s.
	pre post difference	β=1.82, SE=.81, p=.0126*		n.s.
	*group*gesture*time*		β=2.09, SE=1.14, p=.0334*	
Float vs. Glide	*pre-test (time1)*	β=1.17, SE=.57, p=.0199*		β=1.35, SE=.56, p=.0083*
	post-test (time2)	β=2.78, SE=.58, p=<.0001*		β=2.65, SE=.56, p=.0523*
	pre post difference	β=1.6, SE=.80, p=.0232*		β=1.29, SE=.79, p=.0543*
	*group*gesture*time*		n.s.	

Table 18e. The estimates, standard errors and *p* values from the planned complex comparisons for ‘pitch variability’. Blue columns represent significant values. Of particular interest are the group*gesture*time interactions, which would indicate that the differences in pitch variability between the two gestures of each pair are enhanced from pre to post-test at a greater degree for the motor practice group compared to visual practice group.

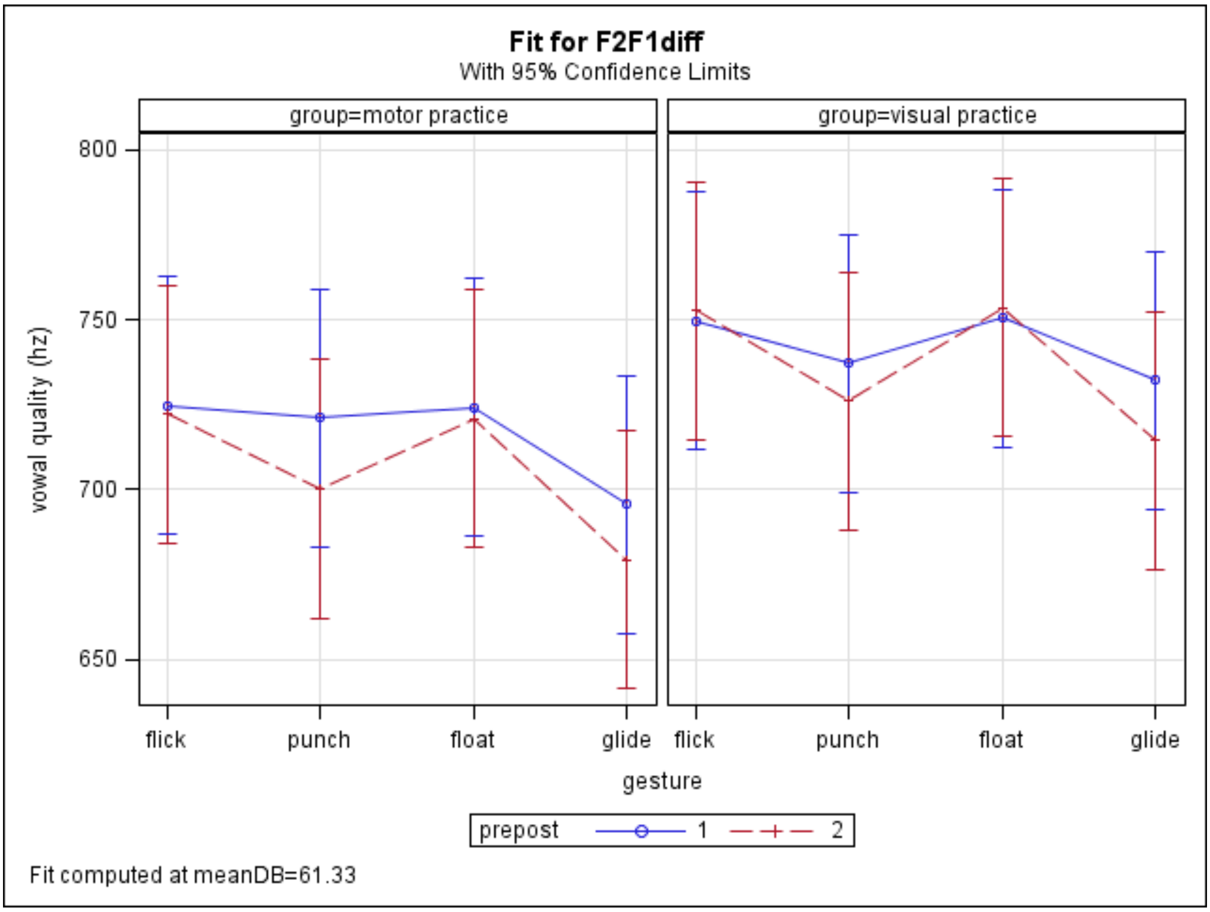


Figure 18f. The model output for "vowel quality – F2 F1 difference".

		Motor Practice	*Visual Practice*
Flick vs. Punch	*pre-test (time1)*	*n.s.*	*β*=12.2, *SE*=7.36, *p*=.0476*
	post-test (time2)	*β*=21.9, *SE*=8.07, *p*=.0033*	*β*=26.4, *SE*=7.44, *p*=.0002*
	pre post difference	*β*=18.1, *SE*=10.3, *p*=.0398*	*n.s.*
	*group*gesture*time*	*n.s.*	
Float vs. Glide	*pre-test (time1)*	*β*=28.5, *SE*=7.31, *p*=<.0001*	*β*=18.3, *SE*=7.2, *p*=.0054*
	post-test (time2)	*β*=41.6, *SE*=7.5, *p*=<.0001*	*β*=39.01, *SE*=7.2, *p*=<.0001*
	pre post difference	*n.s.*	*β*=20.6, *SE*=10.1, *p*=.0215*
	*group*gesture*time*	*n.s.*	

Table 18f. The estimates, standard errors and *p* values from the planned complex comparisons for 'vowel quality. Blue columns represent significant values. Of particular interest are the group*gesture*time interactions, which would indicate that the differences in vowel quality between the two gestures of each pair are enhanced from pre to post-test at a greater degree for the motor practice group compared to visual practice group.

Acoustic analyses showed performance improvements for both motor and visual practice at different levels, and the gains were significantly higher for motor practice. The performance gains were observed after motor practice for all types of acoustic measures except for fundamental frequency for flick-punch pair, and vowel quality for float-glide pair. After the visual practice, on the other hand, performance gains were observed only for flick-punch pair, and only in terms of duration, mean intensity, and intensity variation. Motor practice resulted in more pronounced gains over visual practice for all of the acoustic measures, except for vowel quality. When compared to visual practice, motor practice led to greater mean intensity and pitch variability differences between flick and punch; and greater duration, mean intensity, intensity variation, and F0 differences between float and glide.

Part 2

Another series of linear mixed-effects models were run for each gesture separately to explore (1) whether motor and visual practice were sufficient on their own to elicit the desired enhancements or abatements for the six acoustic features from pre-test to post-test for each gesture individually, and (2) whether motor practice was more influential than the visual practice in eliciting such desired gains in performance. For each gesture type, six linear mixed models were constructed, and they included a between-subjects fixed factor of group (i.e., motor practice group vs. visual practice group) and a within-subjects fixed factor of time (pre-test scores vs. post test scores). For each subject, random intercepts were added to allow accounting for the different baseline values for each subject. The formulas modelling six dependent measures as a function of time and group are as follows:

(1) duration ~ time + group + (1|subject) + ε
(2) meanDB ~ time + group + (1|subject) + ε
(3) sdDB ~ time + meanDB + group + (1|subject) + ε
(4) F0 ~ time + meanDB + group + (1|subject) + ε
(5) sdPitch ~ time + meanDB + group + (1|subject) + ε
(6) F2-F1 ~ time + meanDB + group + (1|subject) + ε

In order to assess hypothesis (1) we calculated simple slopes of the experimental and control groups as a function of time (pre-test vs. post-test) and tested them for significance. In order to assess hypothesis (2) we calculated the difference in the two slopes and tested them for significance (would be equal to the interaction term between group and time). This would entail a more pronounced difference in the acoustic variables at the post-test compared to pre-test (hence enhanced visual-to-auditory coupling) for the motor practice group compared to visual practice group.

We had hypothesized that for each gesture specific qualities would be enhanced and specific qualities would be abated, which we specify as representative of “better performance”. Figure 19 a-d represents six dependent measures as a function of time for flick, punch, float and glide respectively. The specific directions of the changes that characterize better performance are summarized in Table 19 a-d along with the estimates, standard deviations and *p* values from the simple slopes analyses.

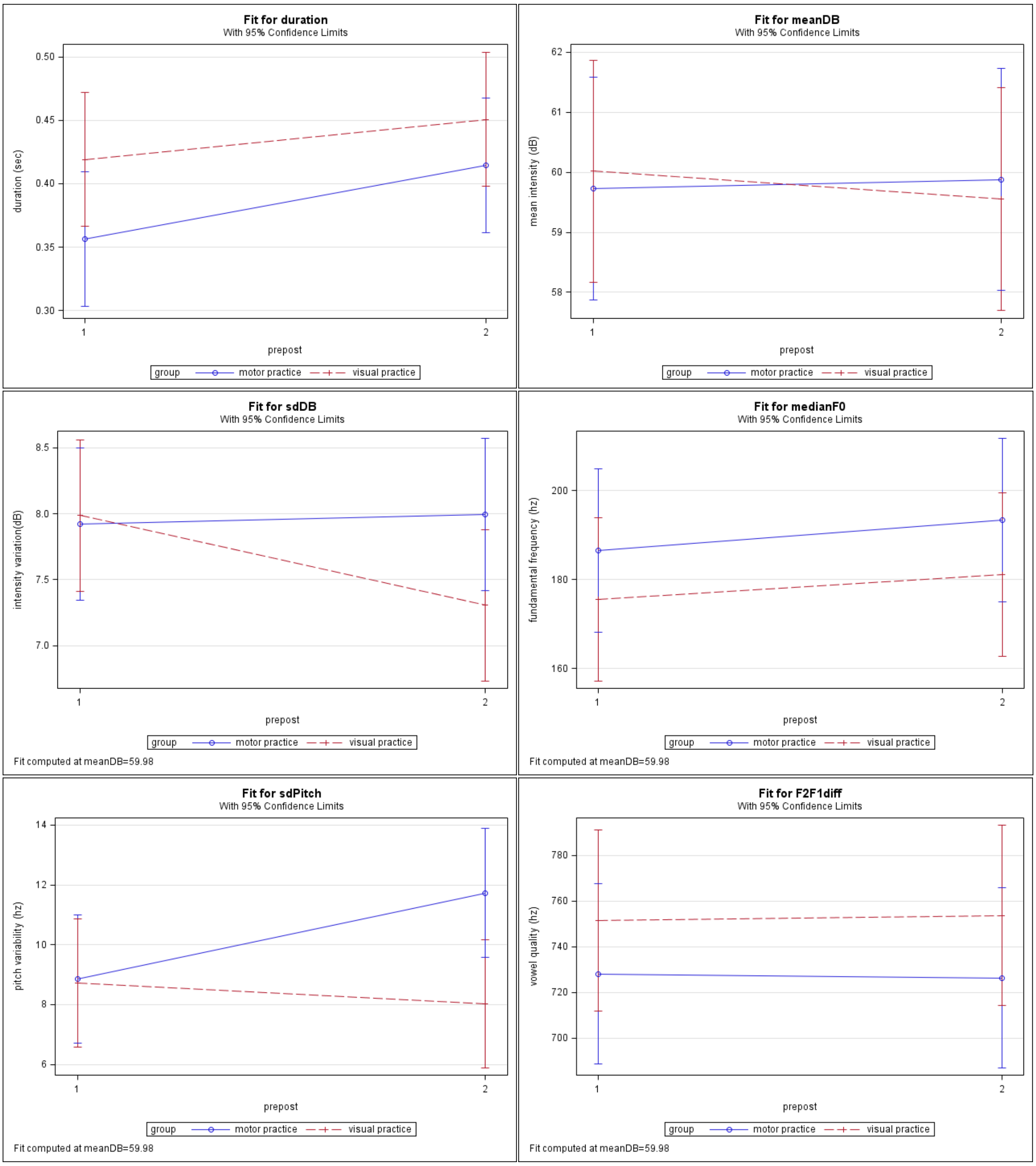


Figure 19a. Six dependent measures as a function of time for FLICK gesture. Error bars represent confidence limits. These are computed as the normal (Wald) confidence limits for the linear predictor.

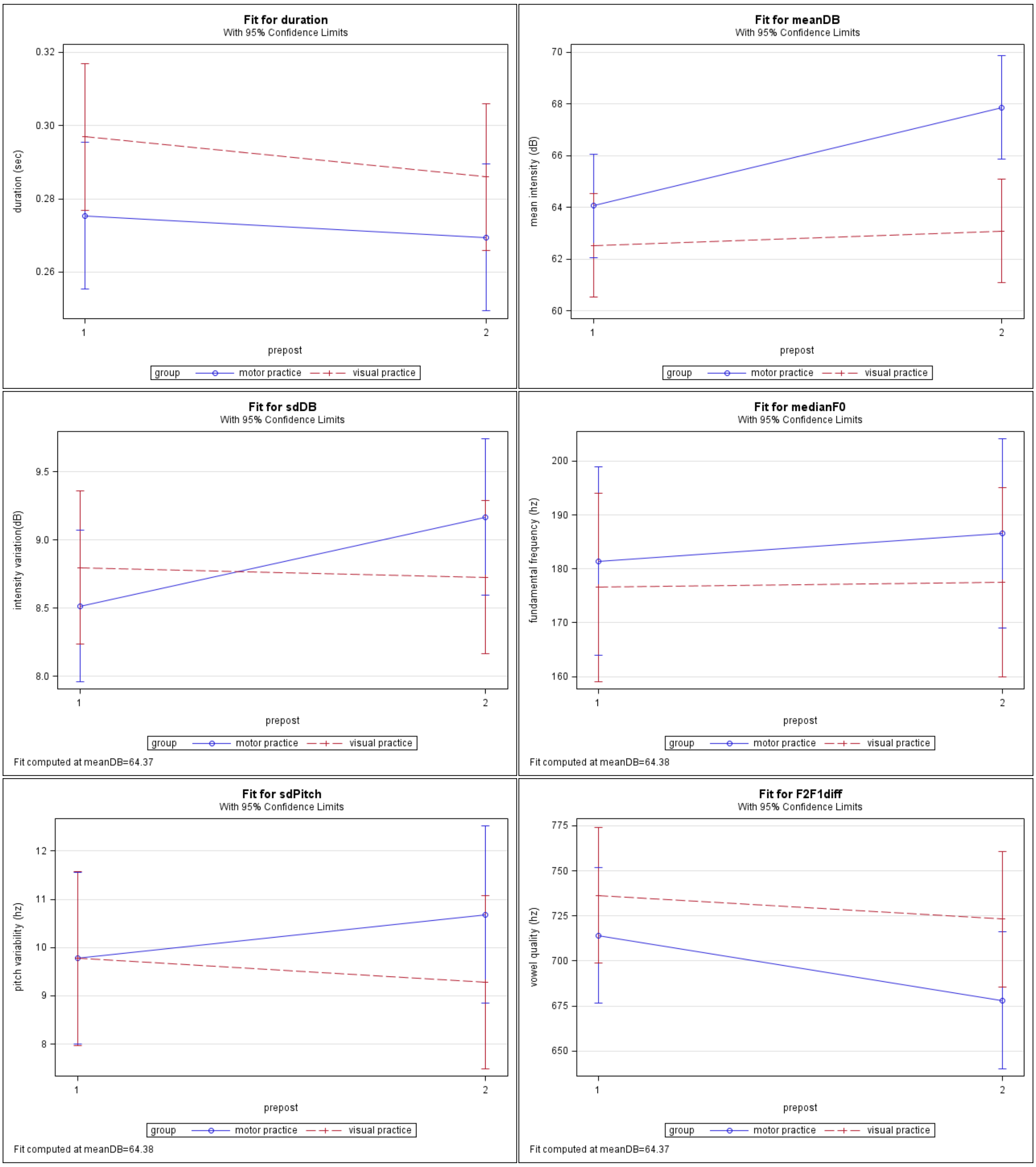


Figure 19b. Six dependent measures as a function of time for PUNCH gesture. Error bars represent confidence limits. These are computed as the normal (Wald) confidence limits for the linear predictor.

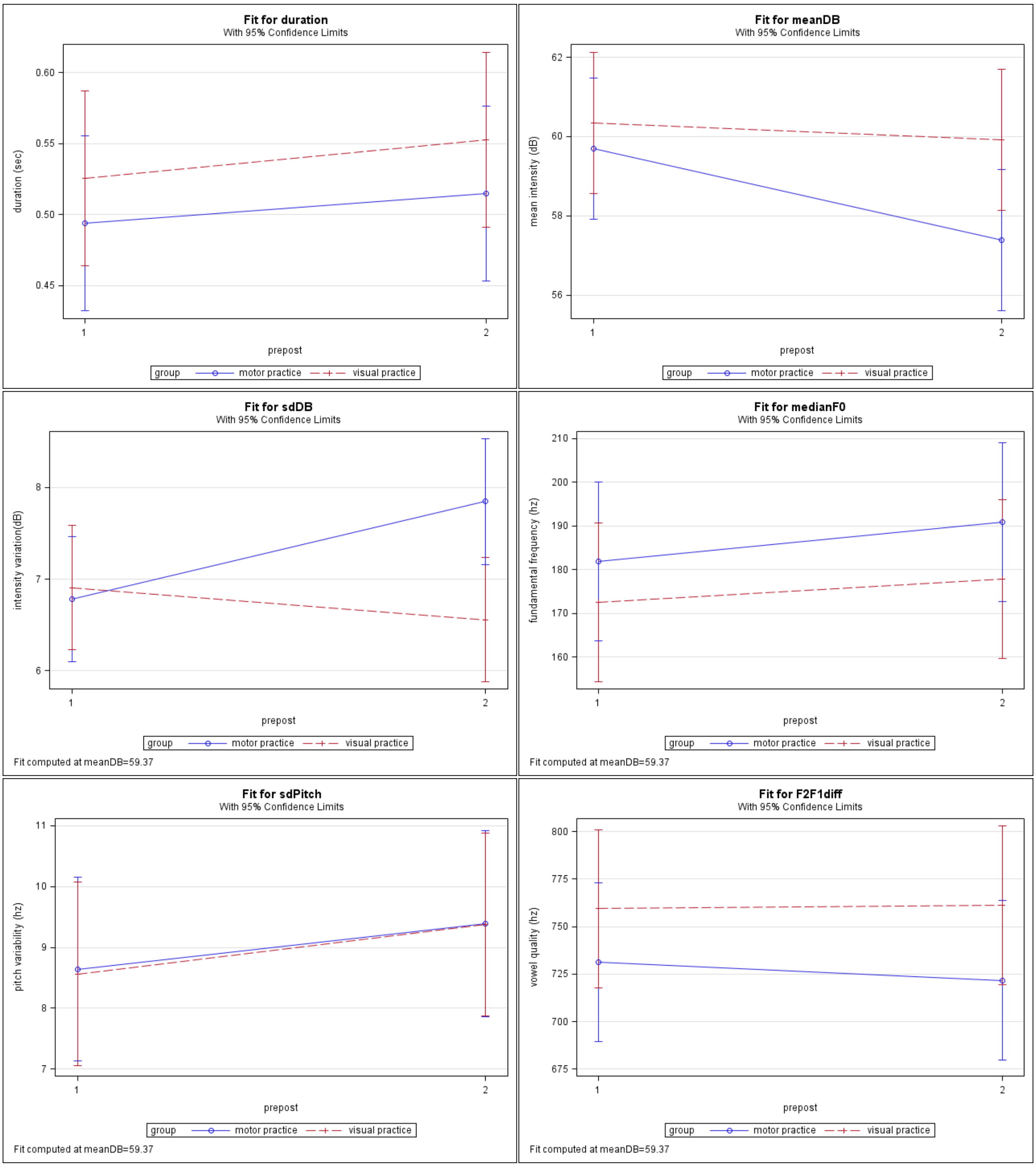


Figure 19c. Six dependent measures as a function of time for FLOAT gesture. Error bars represent confidence limits. These are computed as the normal (Wald) confidence limits for the linear predictor.

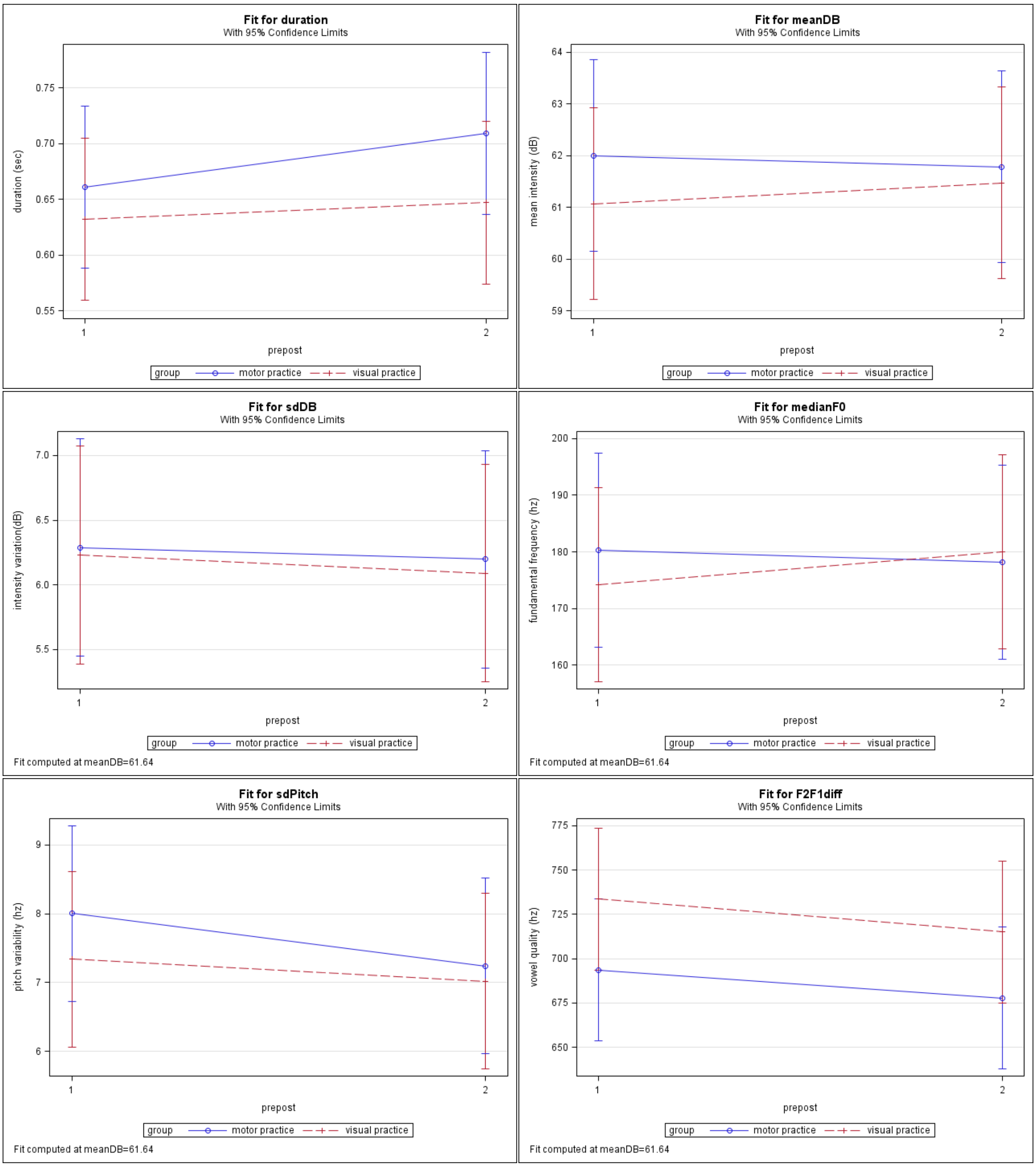


Figure 19d. Six dependent measures as a function of time for GLIDE gesture. Error bars represent confidence limits. These are computed as the normal (Wald) confidence limits for the linear predictor.

FLICK	duration	mean intensity	intensity variation	fundamental frequency	pitch variation	vowel quality (F2-F1)
motor practice	↑ *β*=-.05, *SE*=.008, *p*=<.0001*	↓ *n.s.*	↓ *n.s.*	↑ *β*=-6.9, *SE*=1.22, *p*=<.0001*	↑ *β*=-2.8, *SE*=.66, *p*=<.0001*	↑ *n.s.*
visual practice	↑ *β*=-.03, *SE*=.008, *p*=<.0001*	↓ *n.s.*	↓ *β*=.68, *SE*=.16, *p*=<.0001*	↑ *β*=-5.7, *SE*=1.21, *p*=<.0001*	↑ *n.s.*	↑ *n.s.*
*motor > visual (group*time)*	*F(1,504)=4.64 p*=.0158*	*n.s.*	*F(1,503)=10.83, p*=.0005*	*n.s.*	*F(1,502)=14.19 p*=.0001*	*n.s.*

PUNCH	duration	mean intensity	intensity variation	fundamental frequency	pitch variation	vowel quality (F2-F1)
motor practice	↓ *β*=.005, *SE*=.004, *p*=.0788	↑ *β*=-3.8, *SE*=.28, *p*=<.0001*	↑ *β*=-.65, *SE*=.17, *p*=.0001*	↓ *n.s.*	↓ *n.s.*	↓ *β*=36.1, *SE*=6.87, *p*=<.0001*
visual practice	↓ *β*=.01, *SE*=.004, *p*=.0047*	↑ *β*=-.55 *SE*=.29, *p*=.0295*	↑ *n.s.*	↓ *n.s.*	↓ *n.s.*	↓ *β*=13.3, *SE*=6.09 *p*=.0146*
*motor > visual (group*time)*	*n.s.*	*F(1,500)=62.19, p*=<.0001*	*F(1,499)=9.41, p*=.0101*	*F(1,498)=3.88, p*=.0495*	*F(1,498)=3.62, p*=.0578	*F(1,499)=6.38, p*=.009*

FLOAT	duration	mean intensity	intensity variation	fundamental frequency	pitch variation	vowel quality (F2-F1)
motor practice	↑ *β*=-.01, *SE*=.008, *p*=.0199*	↓ *β*=2.3, *SE*=.26, *p*=<.0001*	↑ *β*=-.06, *SE*=.16, *p*=<.0001*	↑ *β*=-9.02, *SE*=1.08, *p*=<.0001*	↑ *β*=-.74, *SE*=.45, *p*=.0513	↑ *n.s.*
visual practice	↑ *β*=-.02, *SE*=.008, *p*=.0007*	↓ *β*=.42, *SE*=.26, *p*=.0532	↑ *n.s.*	↑ *β*=-5.4, *SE*=1.0, *p*=<.0001*	↑ *β*=-.81, *SE*=.43, *p*=.0292	↑ *n.s.*
*motor > visual (group*time)*	*n.s.*	*F(1,509)=24.78, p*=<.0001*	*F(1,508)=39.07, p*=<.0001*	*F(1,508)=6.16, p*=.0067*	*n.s.*	*n.s.*

GLIDE	duration	mean intensity	intensity variation	fundamental frequency	pitch variation	vowel quality (F2-F1)
motor practice	↑ *β*=-.04, *SE*=.01, *p*=<.0001*	↑ *n.s.*	↓ *n.s.*	↓ *β*=2.15, *SE*=1.5, *p*=.0790	↓ *β*=.75, *SE*=.41, *p*=.0344*	↓ *β*=15.8, *SE*=5.71, *p*=.0029*
visual practice	↑ *β*=-.01, *SE*=.009, *p*=.0673	↑ *n.s.*	↓ *n.s.*	↓ *n.s.*	↓ *n.s.*	↓ *β*=18.4, *SE*=5.7, *p*=.0006*
*motor > visual (group*time)*	*F(1,509)=24.78, p*=.0087*	*n.s.*	*n.s.*	*F(1,505)=13.81, p*=.0001*	*n.s.*	*n.s.*

Table 19 a-d The specific directions of the changes that characterize better performance, along with the estimates, standard deviations and *p* values from the simple slopes analyses. The first row summarizes simple slopes of motor practice as a function of pre-test versus post-test (whether the gain is significant from pre-test to post-test). The second row summarized simple slopes of visual practice as a function of pre-test versus post-test. Third row summarizes group by time interactions, which would entail better performance after motor practice compared to visual practice. Blue columns represent significant results: specific enhancement/abatement predictions that were achieved. Red column indicates that the effect happened in the opposite direction of what was predicted.

When each gesture was investigated separately, simple slopes analyses showed that motor practice led to a flick /da/ that is longer in duration, higher in F0 and pitch variation; a punch /da/ that is higher in mean intensity and intensity variation, and lower in F2-F1 difference; a float /da/ that is higher in duration, intensity variation, F0 and pitch variation and lower in mean intensity; and a glide /da/ that is longer in duration, and lower in pitch variation and F2-F1 difference; all of which were indicative of sounds that are better representative of the gestures being observed.

Visual practice led to better cross-modal couplings in terms of some of the features as well. As shown by the second set of simple slopes analyses, visual practice resulted in a flick /da/ that is longer in duration and higher in F0, a punch /da/ that is lower in duration and F2-F1 difference, and higher in intensity; a float /da/ that is higher in duration, F0 and pitch variation; and glide /da/ that is lower in F2-F1 difference.

However, even when performance enhancements were observed, they were at a lesser extent after passive viewing of the gestures than after active motor practice. As shown by significant simple slope differences, flick /da/s were longer and more variable in pitch following motor practice than following visual practice. Punch /da/s were stronger, more variable in intensity, with a vowel quality that is more to the back and open (low longue) following motor practice than following visual practice. Float /da/s were softer, more variable in intensity, and higher in F0 following motor practice than following visual practice. And glide /da/s were longer and higher in F0 following motor practice than following visual practice.

CHAPTER IV

DISCUSSION

Experiment 1. Cross-modal mapping of observed gestures onto vocal sounds

The purpose of the first experiment was to explore whether there was a systematic relationship between four different hand gestures performed by an expert conductor, and accompanying vocal sounds produced by college students with no significant amount of music background. The participants watched 4 different conducting gestures while they simultaneously vocalized the syllable /da/ in a way that they thought fit the visual gestures.

When the judges listened to the four types of sounds, they could categorize them into their respective gesture type reliably by ear, with very high accuracy ratings that ranged between %85-%100. This showed that there were clear perceptual acoustic markers in each of the sound category which led them to be perceptually distinct. Further acoustic analyses on the responses showed that the participants reliably varied the way they produced /da/ sounds by modifying their acoustic parameters in a predictive way as to match the movement features of the gestures. Specifically they modified their (1) duration to match the duration of the gesture, (2) mean intensity to match the mean/initial velocity of the gesture, (3) intensity variation to match the velocity variation of the gesture, (4) fundamental frequency to match the y-axis elevation of the gesture (only for float-glide pair) (5) pitch variation to match the vertical displacement/velocity of the gesture. The expectations with regard to F0 for the flick-punch pair, and vowel quality did not hold. Lack of sufficient difference in F0 between flick and punch F0s might be at least partly due to punch's increased loudness, as past studies have reported an increase in voice F0 as a function of increased loudness, which involuntarily results from greater resistance by the vocal folds to increased airflow (Gramming et al. 1988; Raphael et al. 2007). Gramming et al. (1988) investigated the change in mean F0 accompanying changes in loudness of phonation during reading, and found that F0 increased by between 0.2 and 0.6 semitones per dB equivalent sound level.

However, the results evidenced that naïve participants with no significant music background could reliably map motional features onto sound features. Past research has documented that people can reliably map sound features onto motional features (Eitan & Granot 2006; Kohn & Eitan, 2009; Küssner et al., 2012, Küssner et al., 2014; Thompson & Luck, 2008; Burger et al., 2011), and they can reliably detect links of auditory-visual stimuli (Lipscomb and Kim, 2004; Kohn and Eitan, 2012). This study contributes to the literature by showing that they can also reliably map visual-motor features onto vocal-motor features.

Experiment 2. Effect of instruction on gesture sound coupling

The purpose of the second experiment was to explore whether the ability to map visual features onto acoustic features is based on an automatic processing/mapping or on a deliberate cognitive strategy. We were interested in whether participants would modify their responses "automatically", when they were not specifically instructed to match what they observe to what they vocally produce, instead just told to "say" /da/ along with the observed gestures.

The results of the perceptual task showed that the judges could reliably guess three of the four gestures (flick, punch and float) from the sound files, although the accuracy scores were significantly below of those from the experiment 1, where the participants were explicitly instructed to "match" the gestures to what they say. The accuracy scores dropped from a range of %85-100 to a range of %42.5-55.This shows that with the weak instruction visual-to-auditory mapping is affected negatively; however there was still enough variation in the sound data that a trained ear could pick up. When asked, most of the participants reported that they did not notice differences in their repeated renditions of the /da/ sounds. However, the judges were able to pick up subtle variations when they were present.

The acoustic analyses further portrayed which specific features the participants modified with weak instruction in relation to strong instruction (experiment 1). Their ‘punch’ /da/ was overall shorter, louder, more variable in intensity, and lower in pitch than their ‘flick’ /da/, which was in line with our hypotheses from the cross-modal task. Similarly, their ‘float’ /da/ was overall softer and more variable in pitch then their ‘glide’ /da/, which was again, in line with our hypotheses. However, the differences were not as pronounced as in experiment 1 as shown by significant interaction effects for all of the six acoustic measures. Overall duration and mean intensity were the most salient features to be modified, for which the participants produced reliable differences between both pairs of gestures. They modified pitch related features of F0 and pitch variability only partially, and they did not show any modification of vowel quality. This suggested that people automatically extracted the durational features from the gestures to map onto the duration of their vocalizations, and they automatically extracted the energy/velocity related features to modulate their intensity levels. It seems that, on the other hand, pitch effects require more deliberate, controlled and conscious cognitive processing as those effects were only partially observed.

The participants in this group reliably modified the duration of their responses in a way in line with our expectations following the actual duration of the gestures (although to a lesser degree). Similarly they reliably modified their sound intensity levels following the energy/velocity profiles of the gestures (although to a lesser degree). This suggests that cross-modal integration of time and intensity is (at least partially) the result of an “automatic” perceptual processing. This makes sense, given that duration and intensity are “amodal” features that are not one modality specific, and could be specified in more than one modality.

Amodal features are those that are not peculiar to a single modality (e.g. audition) and those that can be used to identify an aspect of an object/event in more than one modality (e.g. both auditory and visual) (Lewkowicz, Leo & Simion, 2010). This gives us ample opportunity in life to experience cross-modal associations of time and intensity across different modalities.

For example, everyday life is full of temporal synchronies across modalities. Communication involves both an audible speech and facial/gestural motions or expressions produced in synchrony. When objects move and make sounds, the physical duration of the visual movement usually accompanies the sound that the object emits. Similarly, heavy objects (which require more energy) tend to make louder sounds, or increased physiological energy causes speech to be louder and so forth. On the other hand, pitch is peculiar to the auditory domain, and the cross-modal association of pitch effects with spatial effects is observed less frequently.

This pattern is also in parallel to the development of intersensory association capacities, where babies are shown to be able to match auditory and visual attributes based on temporal synchrony as newborns (Lewkowicz, Leo & Simion, 2010), intensity at 3 weeks (Lewkowicz & Turkewitz, 1980), and pitch at 4 months (Walker et al., 2010; Dolscheid et al., 2012). That intersensory association capacities of time and intensity emerge earliest in life suggests either an innate or at least more easily learned cross modal links of temporal and intensity features. Amodal temporal and intensity related features are also more prominent in everyday life leading the way for stronger and more automatic responses to temporal and dynamic changes cross-modally. In sum, gesture-sound coupling is not entirely based on a deliberate matching strategy, and this coupling is at least partially the result of an automatic perceptual processing and mapping of the visual/motor features onto the acoustic features.

Experiment 3. Role of musical expertise in gesture sound coupling

The purpose of experiment 3 was to explore whether cross-modal association of gesture and sound strengthens as a function of musical expertise. For this reason we compared the gesture-sound mapping ability of musicians (experiment 3) with those of nonmusicians (experiment 1).

During the perceptual task, the judges' responses did not differ when they listened to the sound samples of nonmusicians or musicians. The categorization accuracies were very high for both groups with a range of %85-%100 for nonmusicians, and %77.5-%100 for musicians.

The acoustic analyses, on the other hand, showed subtle differences between musicians and nonmusicians. The differences, when they occurred indicated stronger modifications of acoustic features in the case of musicians. This suggested that musicians' acoustic representations of visual features are more consistent and accurate than those of nonmusicians. Specifically, musicians produced more pronounced differences in duration for float-glide pair; more pronounced differences in mean intensity for flick-punch and float-glide pairs, more pronounced differences in intensity variation for float-glide pair; and more pronounced differences in fundamental frequency for flick-punch pair. Moreover, musicians displayed fundamental frequency differences between flick and punch gestures, which was not previously observed in the control group. The most salient enhancement effect was in intensity where both pairs of gestures were reliably modified at a greater extent in the case of musicians; however enhancements were observed for all of the acoustic features except pitch variation and vowel quality. As unexpected musicians showed no differences in pitch variation for flick-punch pair (contrary to the control group). This finding could be explained by their punch /da/ that was produced with greater loudness than in experiment 1, as increased loudness also results in an increase in voice F0, which involuntarily results from greater resistance by the vocal folds to increased airflow (Gramming et al. 1988; Raphael et al. 2007).

The results showed that musicians and nonmusicians relate acoustic and motional features in very similar ways, however at least some cross-modal effects strengthens as a function of musical background. This study adds to the current literature by validating the past research, which showed that musical training leads to more consistent mappings between audio-visual features. Eitan and Granot (2006), asked university students, some of whom had at least 7 years of music training, to imagine/visualize an animated cartoon character move to several melodic soundtracks, and to mark the character's motion in a forced choice questionnaire incorporating several movement features such as movement type (e.g. walking, running, jumping etc.), direction of movement (e.g. vertical, horizontal, sagittal), and overall energy level. They have noted stronger music-motion associations for musicians for the associations of pitch contour with verticality and laterality (e.g. ascending motion and/or motion to the right for rising pitch contours), and for the associations of IOI (inter-onset-interval) with speed (e.g. accelerating motion for acceleration in the musical phrase). Vertical and lateral associations of pitch height were possibly due to musicians' involvement with musical notation and keyboard, where high pitches are positioned above and to the right of lower pitches (Eitan & Granot, 2006). They found no training related differences with regard to dynamics, a finding that they explain by suggesting spatio-kinetic associations of loudness are determined by their everyday connotations.

Kussner et al. (2014) reached similar conclusions by asking musically trained and untrained participants to represent several sound stimuli differing in pitch, intensity and tempo with their right hand. When compared with musically untrained participants, the musically trained showed larger positive correlations between pitch and verticality (e.g. higher pitch for higher elevation) as in line with Eitan & Granot, 2006. However as different from Eitan & Granot (2006) they have also found enhancement effects for the link between loudness and movement along z axis (e.g. increased loudness for moving forward), and between loudness and muscular energy (e.g. increased loudness for more hand shaking events).

The current study reaches similar conclusions by reversing the paradigm, where participants are asked to produce sound with different types of gestures. When compared to musically untrained (experiment 1), musically trained participants (experiment 3) showed stronger manipulations of duration (in response to duration changes), stronger manipulations of mean intensity (in response to energy/velocity changes), as well as stronger manipulations of intensity variation (in response to velocity variation) and fundamental frequency (in response to vertical elevation).

Most musicians, such as the ones in our study, have considerable amount of experience with conductors either in orchestral or choir settings, where they have to follow the lead of a conductor to modulate their vocalizations or the sounds of their instruments, and have to match the information provided through gestures to what they produce aurally. This experience might have helped them to form stronger and more consistent mappings between motional and sound features. It is also possible that musical practice, in particular the motional experience of playing an instrument, and hearing the accompanying sounds, plays a role in sharpening the spatial and kinetic associations with musical parameters.

Despite the enhancement effects, it is important to note that the kinds of mappings remained consistent across most Western individuals regardless of musical background. In sum, the results suggested sound-motion mappings, in general, do not stem from musical experience but from more general sources as shown by the advanced performance level of nonmusicians.

Experiment 4. Use of point light displays in gesture sound coupling

The purpose of experiment 4 was to explore the role played by full body information versus spatio-kinematic information in the cross-modal mapping of gesture onto sound. This was achieved by presenting the participants with motion-capture based dynamic point light patterns (PLD) of the four gestures to eliminate the featural information from the movements.

During the perceptual task, the judges' responses did not differ when they listened to the sound samples of participants watching point light representations (experiment 1) or full body representations (experiment 4). The categorization accuracies were very high for both groups with a range of %85-%100 for experiment 1, and %90-%100 for experiment 4.

The acoustic analyses portrayed which specific features the participants modified with point light displays. They have shown all the modifications as hypothesized except for intensity variation differences for flick-punch, and fundamental frequency differences for float-glide which were not achieved in the PLD group. This suggested that even when featural body information was missing the motional-sound links could still be reliably formed based on the spatio-kinematic information alone. Basically positional and velocity information is sufficient and full body representation is not necessary to successfully map duration, energy and spatial information onto duration, intensity and pitch. This result is consistent with previous literature which showed that point light displays (originally described by Johansson, 1973) are sufficient to convey e.g. gait of friends (Dittrich, Churchill, & Weidenbacher, 1994), gender of a walking person (Kozlowski & Cutting 1977; Mather & Murdoch 1994), the type of human action (Dittrich, 1993) and basic emotions portrayed by body movements (Atkinson, Dittrich, Gemmel and Young, 2004; Brownlow, Doxin, & Radcliffe, 1997; Dittrich, Troscianko, & Morgan, 1996).

However, the analyses also showed some discrepancies between the two groups (control and PLD) with regard to the strength of the associations. Participants watching PLDs produced shorter flick /da/s than the control group; however this effect turned out to be a spurious effect caused by the absence of finger markers during the recording session of flick gesture in motion capture. In order to gather positional data accurately from the movements, the joint markers needed to be placed at a clear distance from one another. In our pilot testing where finger markers were attached the data could not be recorded reliably, hence the final part of the flick action where fingers were closing (marking the tail of the gesture) was not present in the point light renditions. The absence of this very last part of the gesture resulted in the perception of a shorter gesture than it really is.

There was no discrepancy between the two groups with regard to the effect of mean intensity. Hence, the absence of duration and mean intensity differences between the two groups suggested that cross-modal integration of (relatively simple) temporal and intensity features could be formed as reliably even when full body information was missing. Participants watching PLDs were as accurate and consistent as participants watching full body representations for extracting duration and energy to map onto duration and intensity. The discrepancies were present for intensity variation, fundamental frequency and vowel quality. For these features, the modifications were more pronounced for the control group when integrating velocity variation with intensity variation for flick-punch pair; when integrating F0 with verticality for float-glide pair, and when integrating vowel quality with motional features for float-glide pair. On the other hand, the modifications were more pronounced for PLD group when integrating velocity variation with intensity variation for float-glide pair; and when integrating F0 with verticality for flick-punch pair.

Experiment 5. Role of self-produced auditory feedback in gesture sound coupling

The purpose of the experiment 5 was to explore the role played by self-produced auditory feedback in the cross-modal mapping of gesture and sound. This was achieved by masking auditory feedback, so that participants could rely solely on their motor representations and kinaesthetic feedback from the movement of their vocal apparatus during the cross-modal task.

The judges could reliably identify which gesture the participants were observing from the sound data whether they had access to their auditory feedback (experiment 1) or they were deprived of their auditory feedback (experiment 5). The categorization accuracies were very high for both groups with a range of %85-%100 for control participants, and %87.5-%97.5 for auditory feedback deprived participants. This was an interesting finding, showing that the participants could produce comparable levels of acoustic variations in their /da/ sounds with or without access to the auditory component of their vocal output.

The acoustic analyses further portrayed that participants in this group reliably modified all of the acoustic features as hypothesized except for fundamental frequency for both pairs of gestures of interest, and pitch variation and vowel quality for flick-punch pair only. Duration and energy/velocity from the movements were not only reliably mapped onto duration and intensity in the accompanying sound counterparts, but the effects were even "enhanced" when auditory feedback was absent, which was an unexpected finding. Basically, the participants with masked auditory feedback produced more pronounced differences between the flick and punch in terms of duration, and more pronounced differences between float and glide in terms of duration, mean intensity, and intensity variation. On the other hand, pitch related effects were either partially absent, or less pronounced when auditory feedback was absent. Specifically no fundamental frequency changes observed for any of the pairs, and pitch variation changes were not present for the flick-punch pair. Formant frequency changes were either absent (flick-punch) or at the same level (float-glide) with those of control group.

This result is consistent with the literature evidencing that lack of auditory feedback leads to deterioration of intonation accuracy and fine control of F0 while singing (Elliot &

Niemoeller, 1970; SchultzCoulon, 1978; Murbe, Pabst, Hofmann, & Sundberg, 2002; Ward & Burns, 1978; Erdemir & Rieser, 2016). Moreover, studies have shown that when pitch information is suddenly raised or lowered during speech and singing, participants would adapt to the perturbation by changing the pitch of their speech in the direction opposite of the shift (Kawahara, 1995; Burnett et al., 1998; Jones and Munhall 2002; Jones and Keough, 2008), a consistent finding which has been interpreted as support for the idea that F0 control is reliant on auditory feedback. It seems like auditory feedback plays an important role in fine control of pitch, where auditory sensory feedback provides necessary information to form speech motor goals and for error correction when the target pitch does not match intended pitch. Such process would allow for updating the internal representation (see Flanagan and Wing, 1993 for internal models) of the mapping between pitch output and the motor system that controls its production.

The current experiment adds to the existing literature by showing that auditory feedback is also important for cross modal integration of spatial information onto pitch. Given that pitch accuracy highly depends on auditory feedback utilization, it makes sense that cross modal mapping of pitch would depend on auditory feedback. This suggests that association of verticality and pitch portrays mapping of visual gesture onto the acoustic sound where auditory feedback plays an important role. On the other hand, cross modal integration of time and energy cues onto time and intensity cues portrays a visual-to-vocal-motor mapping without the intervening effect of auditory feedback. People seem to directly map visual information of time and energy/velocity onto the vocal-motor system directly.

This result suggests an automatic integration mechanism that is directly from visual-motor onto the vocal motor system that bypasses the auditory feedback mechanism. The fact that temporal and intensity associations were even enhanced when auditory feedback was absent suggests that motor representations and kinaesthetic feedback play a greater role than the role played by auditory feedback to form sensorimotor integration of temporal and intensity related features. More attention resources being devoted to the motor representations of planned vocal actions (rather than acoustic output itself) might have resulted in this superiority effect. It is notable that reliance on motor representations resulted in superiority effects for the temporal and intensity related features (duration, mean intensity and intensity variation) but not for pitch related features (F0, pitch variation and vowel quality).

Erdemir and Rieser (2006) had shown that singers with significant voice training background relied less on auditory feedback for fine control of pitch when singing; similarly in our study naïve nonmusicians seem to be already skilled at such cross-modal task when the task involved matching temporal and intensity related features across visual and auditory domains. This would entail that cross modal coupling of (relatively simpler) time and intensity related features are more intrinsic than those of pitch related mappings, and that people are already skilled at extracting time and energy/velocity from visual stimuli to map onto temporal and intensity levels, possibly due to direct involvement of the vocal-motor system. This study suggested that another possible reason why cross-modal mapping of temporal and intensity features are easier could be due to the direct mapping of visual properties onto the vocal-motor system by bypassing the auditory feedback loop.

General discussion for experiments 1, 2, 3, 4 & 5

The findings from experiments 1-5 indicated that time and energy/velocity related features from the visual gestures were automatically (experiment 1) transferred into time and intensity levels within the vocal motor system without the intervening effect of auditory feedback (experiment 5). These effects were as strongly present even in the absence of significant experience with musical practice (experiment 1) and in the absence of full body representations (experiment 4) from observed gestures. Pitch related associations, on the other hand, were either absent or not as strongly present with several of our experimental manipulations including the weak instruction, point light display and auditory feedback masking conditions, whereas they were slightly enhanced in the musicians group (F0 for the flick-punch pair). Even the control group (nonmusician college students) lacked fundamental frequency differences for flick-punch pair, and formant frequency differences for both pairs of gestures.

Origin and shaping of cross-modal correspondences

Developmentally infants are born ready to be able to match certain auditory and visual attributes. Research shows that even newborn infants can make at least some across-modal associations intuitively and readily, such as integrating the amount of visual stimulation with the amount of auditory stimulation (Lewkowicz & Turkewitz, 1981; Gardner et al., 1986), matching audio-visual elements of objects based on co-location and synchrony (Morrongiello et al., 1998; Slater et al., 1999), and matching a monkey facial gestures and accompanying vocal sounds based on temporal synchrony (Lewkowicz, Leo & Simion, 2010). Newborn infants later develop their cross-modal integration abilities to be able to match auditory and visual attributes based on intensity (brightness and loudness) at 3 weeks of life (Lewkowicz & Turkewitz, 1980). It is at 4 months, when they can show pitch and height associations such that that looking at the sight of animated ball that rose and fell more, when it is accompanied by a rising and falling pitch pattern rather than opposite (Walker et al., 2010, Dolscheid et al., 2012); and it is at 1 year when they can match tones that rise or fall in frequency with arrows that point up or down in space, respectively.

This pattern along with the results from the current study shows that intersensory association capacities of time and intensity emerge earliest in life, suggesting either an innate or at least more easily learned cross modal associations for temporal and intensity related features. Pitch relations, on the other hand, appear slightly later in life, and they might depend more on a learned aspect of perception. This corroborates the idea that humans begin life with a broadly tuned perceptual-motor system that makes it possible for them to be able to make certain cross-modal associations even at birth and provide the basis for learning other kinds of associations through cortical maturation or experience.

Audio-visual links of temporal synchrony and intensity are more reliably and readily present in everyday life events that we are constantly bombarded with. Pitch and height related links, on the other hand, appear less reliably and consistently. For instance, when mothers teach their infants novel object-word associations they temporally synchronize object motion with the target word, as if they implicitly know the importance of synchronous multimodal stimulation for their babies (Gogate, Bahrick and Watson, 2000). Speech is full of temporal synchrony between facial/gestural movements and acoustic output. The pace of motion is universally coordinated with tempo in dances, marches and all other music-related genres. Similarly speech is full of intensity related mappings between motional and acoustic features. For example, experiencing aroused emotions such as fear, anger or happiness causes high sympathetic arousal providing the body with extra energy, speed and strength (Johnstone & Scherer, 2000; Scherer 2003); hence we speak louder when we have more energy flow in the body. Or when heavier objects drop they tend to make louder sounds, and we know that more energy is required to carry or push a heavier object. In acoustics, changing the energy of the intermitted sound produces dynamic changes. And mathematically there is a positive correlation between velocity and kinetic energy, such that faster speed entails more kinetic energy. One reason why these links appear more readily in everyday life could also be due to the “amodal” nature of time and intensity features, which mean those features are not peculiar to a single modality (e.g. audition) but they can be used to specify objects/event both aurally and visually. This gives us ample opportunity in life to experience cross-modal associations of time and intensity across different modalities.

On the other hand, the link between pitch and verticality appears less straightforward. Western listeners speak of pitch in terms of vertical motional terms of ‘high and low’, ‘rise and fall’. However, this relationship could be culturally and historically rooted rather than being universal and biological in origin. For instance, other cultures have used alternative labels for pitch positions, such as ‘sharp or heavy’ in ancient Greece, ‘small or larger in Bali and Java, or ‘young and old’ in Amazon (Zbikowski, 1998). Moreover, congenitally blind participants seem to completely ignore the association of pitch with spatial elevation when asked to imagine an object moving in a way appropriate to the music (Eitan, Ornoy and Granot, 2012). Ashley (2005) has shown that pitch and spatial contour relationships could be obliterated following some training, suggesting pitch verticality mapping is learned.

Still, pitch height is strongly associated with spatial verticality in Western adults with no significant musical background, and we have no explanation for why higher pitches are positioned above lower pitches in musical notation as a convention. Pitch and verticality represents even an unconscious cognitive processing as pitch increases lead to perception of spatial elevation (Roffler & Butler, 1968). When higher pitches correspond with a lower visual signal, brain responses that start earlier than the onset of auditory stimulus are detected signaling incongruency (Widmann et al., 2004). Also, four months of life is still early to show pitch height and verticality associations (Walker et al., 2010). Moreover, consider cartoons with iconic clichés for motion. Consider Wile E. Coyote in *Road Runner* as he jumps repeatedly to reach the road runner at the corner of a cliff, while the upward movement is simultaneously accompanied by a rising pitch. Or consider Baby Herman in the *Who Framed Roger Rabbit*, as he skims through the refrigerator from bottom to top, and the upward movement is accompanied by a rising pitch, which is also in line with the Western tradition of notation. Similarly, from experience we know that singers usually employ a low head position during the singing of low pitch notes, and high head position during the singing of high pitch notes while opening their mouth wider. Therefore, it is less clear whether pitch and verticality appears as a universal and innate cross modal link or a link that is learned based on cultural norms. However, it seems like cross modal links of temporal and intensity features are more easily assessed, less prone to manipulations and, perhaps more biologically based.

In sum, the results from first 5 studies seem to provide support for the idea that temporal and intensity mappings rely more strongly on an automatic processing and a direct mapping of visual information onto vocal-motor system, which might has its roots biologically. On the other hand, pitch related associations might rely, at least partially, on learned and culturally specified conventions as determined by everyday connotations. Such view should still be able to explain whether those conventions are based on arbitrary agreements, or are still based on any evolutionary or biological roots.

Experiment 6. Effect of gestural motor practice in gesture sound coupling

The purpose of experiment 6 was to probe the underlying mechanism mediating the coupling of visual gestures and acoustic sounds, and to explore whether motor practice of observed gestures enhances visual-to-auditory coupling. For this reason the participants actively practiced the visual gestures by imitating them several times, whereas the control group passively observed the gestures for the same amount of time. We had predicted that active motor practice would help participants bind visual features with accompanying acoustic ones, and that they would show increased improvement following the motor practice.

The perceptual task was designed to reliably predict /da/ sounds that were better representative of their respective gestures by comparing pre-test and post-test renditions. The judges reliably marked post-test /da/ sounds as better sounding for the motor practice group for flick (%75 of the time), punch (%90 of the time), float (%80 of the time) and glide (%66 of the time). On the other hand, visual practice led to performance increments only for flick with percent accuracy of %68, and for float with a percent accuracy of %66. Therefore, perceptually the /da/ sounds were better representative of their respective gestures following the motor practice group than following visual practice.

Acoustic analyses showed performance improvements for both motor and visual practice at different levels, and the gains were significantly higher after the motor practice. The performance gains were observed after motor practice for all types of acoustic measures except for fundamental frequency for flick-punch pair, and vowel quality for float-glide pair. After the visual practice, on the other hand, performance gains were observed only for flick-punch pair, and only in terms of duration, mean intensity, and intensity variation. Motor practice resulted in more pronounced gains over visual practice for all of the acoustic measures, except for vowel quality. The strongest gain was for mean intensity where all four gestures were modified at a greater extent in the post-test following motor practice. Specifically when compared to visual practice, motor practice led to greater mean intensity and pitch variability differences between flick and punch; and greater duration, mean intensity, intensity variation, and F0 differences between float and glide.

When each gesture was investigated separately, motor practice led to a flick /da/ that is longer in duration, higher in F0 and pitch variation; a punch /da/ that is higher in mean intensity and intensity variation, and lower in F2-F1 difference; a float /da/ that is higher in duration, intensity variation, F0 and pitch variation and lower in mean intensity; and a glide /da/ that is longer in duration, and lower in pitch variation and F2-F1 difference; all of which were indicative of sounds that are better representative of the gestures being observed.

Visual practice led to better cross-modal couplings in terms of some of the features as well, although they were at a lesser extent. Basically visual practice resulted in a flick /da/ that is longer in duration and higher in F0, a punch /da/ that is lower in duration and F2-F1 difference, and higher in intensity; a float /da/ that is higher in duration, F0 and pitch variation; and glide /da/ that is lower in F2-F1 difference. Recall that the perceptual judgment scores were not significantly different from chance level for punch and glide. Acoustic analyses were able to pick up subtle differences in the acoustic data that a trained ear could not pick up.

However, even when performance enhancements were observed, they were at a lesser extent after passive viewing of the gestures than after active motor practice. Specifically, flick /da/s were longer and more variable in pitch following motor practice than following visual practice. Punch /da/s were stronger, more variable in intensity, with a vowel quality that is more to the back and open (low longue) following motor practice than following visual practice. Float /da/s were softer, more variable in intensity, and higher in F0 following motor practice than following visual practice. And glide /da/s were longer and higher in F0 following motor practice than following visual practice.

The results from this study, for the first time, shows evidence that overtly activating the motor pathways corresponding to the execution of observed gestures for a short amount of time, helps to strengthen visual-to-auditory mapping. This leads to the assumption that motor representations underlying observed sequences potentially play a role in the cross-modal transfer from vision to sound. There are two possible explanations about why motor practice might have led to increased improvement:

(1) First possible explanation is that active imitation of the gestures activates the motor pathways responsible for the execution of those movements, and the increased cortical excitability helps to link the motional features with acoustic features. This suggests that

increased structural similarity between movement and sound features might be *due to better mapping of the visual representation of observed action onto one's own motor representation through active imitation*. Increased cortical excitability along with stronger motor representations that are internalized by active imitation might have helped participants to bind the motional information in the gestures with acoustic features. In other words, gestural motor system activity might have helped to transfer knowledge from visual-motor system onto the vocal-motor system. It could be that participants improved their performance by practicing the gestures because a) either the visual representations are directly mapped onto their own motor repertoire via motor practice, thereby they fit the motor features onto the acoustic features better; b) and/or there is transfer from the gestural motor system to the vocal-motor system, such that gestural motor activity positively affects vocal motor activity, and thereby structural similarity between the visual and acoustic features.

(2) Another possible explanation is that by actively rehearsing the gestures the participants noticed more of the representative features that were present in the gestures (e.g. "float fells smoother and longer after the motor task"). The motor practice itself might have helped the participants to have a better representation of the gestures and realize gesture specific features at a greater extent, which are then used to map onto the acoustic ones. However, passive visual practice, which allowed participants to realize the visual features even at a greater extent, did not improve the cross-modal mapping at the same level as did motor practice. Therefore it seems more likely that it is a "motor" representation of the features, rather than "visual" that is extracted and used to map onto the acoustic features.

Interestingly, even a brief passive visual observation of the gestures resulted in the enhancement of some cross-modal links. There are two possible explanations for this effect: (1) focusing on the visual representations of the gestures in more detail might have helped participants to process the motional features that differentiate the four gestures at a deeper cognitive level. Thereby, they might have become slightly more consistent and proficient in their vocal responses given in response to the motional features that the participants processed at a deeper cognitive level. A second explanation is that (2) passive observation might have led to the activation motor pathways underlying observed gestures (although to a lesser degree than that of motor practice) which then might have resulted in enhanced sensori-motor integration. Given that cortical excitability is present even with passive observation of simple hand movements (Clark et al., 2003), this is a potential possibility. If the same basic motor pathways are activated during the passive viewing process as in the motor imitation task (although to a lesser degree than in motor imitation), this activation might have helped to extract visuo-motor features to map onto acoustics features.

In sum, active motor practice of the gestures resulted in stronger and more consistent cross-modal associations than passive viewing; and the enhancement effects were present for all six of the acoustic features. This suggests that motor representations play an important role in the cross-modal transfer from visual modality to auditory modality.

Conclusion

The findings from the series of studies presented here extend the existing evidence about cross-modal mappings of sound and motion, as well as the research about sensory-motor integration and musicianship. The results reflect a strong coupling between visual and auditory processes, where the perceptual interpretation of dynamic visual stimuli affects the acoustic output in a predictable way. We have explored the strength and nature of this relationship by manipulating the kinds of information available to the participants during the cross-modal task, and have shown that such coupling is at least partially intrinsic/automatic, especially when the features to be mapped involved time and intensity related features rather than pitch related features.

Participants reliably mapped visual features onto acoustic features, especially those related to time and energy/intensity, even when they did not have any significant amount of musical background (experiment 1), not explicitly told to match what they see to what they say (experiment 2), deprived of featural information from the visual stimuli so that only spatio-kinematic information is remained (experiment 4), and deprived of their own auditory feedback and had to rely on their motor representations of vocalizations (experiment 5). Such cross-modal mapping is (at least partially) mediated by the gestural motor system since active motor practice is shown to enhance cross-modal links unlike passive visual practice, which yielded into only brief enhancements if ever (experiment 6). The findings from these series of studies extend on the existing evidence of cross-modal literature by suggesting a strong and intrinsic coupling between visuo-motor and auditory processes, that does not depend on a deliberate/learned/conscious cognitive strategy (experiment 2), musical background (experiment 3), full body representations (experiment 4), or auditory feedback (experiment 5); and one that is strengthened by the absence of auditory feedback with increased attention to vocal-motor representations (experiment 5), as well as by motor practice of observed movements, which activates the brain regions involved in executing the movements (experiment 6). The important role of the motor system during this cross-modal task extends the current literature by suggesting an intrinsic coupling between the visual and auditory systems that depends on a motor resonance system as a binding factor.

Implications

From a music education perspective, the findings suggest that conductors, musicians, music students could benefit from using every day-life familiar gestures in music teaching/learning settings. Especially children without any prior musical background could better learn music by training the cross-modal links of gesture and sound using active motor engagement. Musical activity involves movement itself, but the emphasis given to expressive gestural body language is limited in most educational settings. When music is thought, it would help the music students if it were thought through movement. The music student could walk and swing their arms, or conduct while they sang or listened to music; which would supposedly help them to form specific cross-modal links and register the sound features at a deeper level.

Another implication is that children and adults with speech-motor related problems could benefit from therapies based on sound-gesture couplings. If adults have natural connections between various motion features & sound features then incorporating these associations into therapies might help clinical populations with poor speech-motor abilities. For example, an intervention method called Auditory-Motor Mapping Training (AMMT), has been successfully shown to promote speech production in children with ASD, directly by training the cross modal association of sounds and their articulator actions using intonation and bimanual motor activities (Wan, Bazen, Baars, Libenson, Zipse, Zuk, & Schlaug, 2011). Moreover, past research has indicated a high rate of co-occurrence between deficits in motor and language skills in children; and a growing body of literature has shown associations between language and music skills, such that musical deficits being related to speech-language deficits such as language impairment (Gordon, Shivers, Wieland, Kotz, Yoder, & McAuley, 2015; Gordon, Jacobs, Schuele, & McAuley, 2015) and stuttering (Falk, Muller & Dalla Bella, 2015; Etchell, Ryan, Martin, Johnson & Sowman, 2016; Etchell, Johnson, & Sowman, 2014; Weilend McAuley, Dilley & Chang, 2015). Therefore, musical practice in general, where auditory and motor systems are co-activated; or practices that specifically target production of natural gesture-sound associations by using active motor engagement could be used as an effective way to improve speech-motor problems in clinical populations.

REFERENCES

Alluri, V., Toiviainen, P., Jaaskelainen, I. P, Glerean E., Sams, M. & Brattico, E. (2012) Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm, *Neuroimage, 59*, 3677-3689.

Atkinson, A. P., Dittrich, W. H., Gemmell, A. J., & Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, *33*(6), 717-746.

Aschersleben, G. (2002). Temporal control of movements in sensorimotor synchronization. *Brain and Cognition, 48*, 66-79.

Ashley, R. (2005). Cross-modal processing of melodies: Musical, visual, and spatial aspects. Manuscript submitted for publication.

Aziz-Zadeh L, Maeda F, Zaidel E, Mazziotta J, & Iacoboni M.(2002). Lateralization in motor facilitation during action observation: a TMS study. *Experimental Brain Research, 144*, 127–31

Boersma, P., Weenink, D., (n.d). Praat Speech Processing Software, Institute of Phonetics Sciences of the University of Amsterdam. http://www.praat.org

Bosbach S, Cole J, Prinz W, & Knoblich G. (2005). Inferring another's expectation from action: the role of peripheral sensation. *Nature Neuroscience 8*:1295–9

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., Porro, C. A. & Rizzolatti, G. (2004) Neural circuits involved in the recognition of actions performed by non-conspecifics: an fMRI study. *Journal of Cognitive Neuroscience. 16*,114–126.

Burger, B., Thompson, M.R., Saarikallio, S., Luck, G., & Toiviainen, P. (2011). Influence of musical features on characteristics of music-induced movements. *In Proceedings of the 11th International Conference on Music Perception and Cognition*. pp. 425-428.

Burnett, T. A., Freedland, M. B., Larson, C. R. & Hain, T. C. (1998) Voice F0 responses to manipulations in pitch feedback*, Journal of the Acoustical Society of America, 103,* 3153–3161.

Brownlow S, Dixon A R, Egbert C A, Radcliffe R D. (1997) Perception of movement and dancer characteristics from point-light displays of dance. *The Psychological Record 47* 411- 421.

Chen, J. L., Penhune, V. B. & Zatorre, R. J. (2008). Listening to musical rhythms recruits motor regions of the brain. *Cerebral Cortex, 18*, 2844–2854.

Clark, S., Tremblay, F., & Ste-Marie, D. (2003). Differential modulation of corticospinal excitability during observation, mental imagery and imitation of hand actions. *Neuropsychologia, 42*, 105–112.

Cohen, J. (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement* 20 (1): 37–46.

Cook, P., Rouse, A., Wilson, M., & Reichmuth, C. (2013). A California sea lion (Zalophus californianus) can keep the beat: motor entrainment to rhythmic auditory stimuli in a non vocal mimic. *Journal of Comparative Psychology*, *127*(4), 412.

Dalla Bella, S., Giguère, J.-F., & Peretz, I. (2007). Singing proficiency in the general population. *The Journal of the Acoustical Society of America, 121*(2), 1182–1189.

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music, 21*,103–113.

Dittrich W H, Churchill A, & Weidenbacher H. (1994) Recognition of friends by different movements under point-light conditions. UH-Psychology-Reports (1.7), University of Hertfordshire, Hatfield, Herts, UK.

Dittrich, W. H., Troscianko, T., Lea, S. E. G., & Morgan, D. (1996). Perception of emotion from dynamic light-point displays represented in dance. *Perception, 25*, 727-738.

Dolscheid, S., Hunnius, S., Casasanto, D., & Majid, A. (2012). The sound of thickness: Prelinguistic infants' associations of space and pitch. In *the 34th Annual Meeting of the Cognitive Science Society (CogSci 2012)* (pp. 306-311).

Elliot, L., & Niemoeller, A. (1970). The role of hearing in controlling voice fundamental frequency. *International Journal of Audiology*, *9*, 47–52.

Eitan, Z. & Granot, R. (2006). How music moves: Musical Parameters and Listeners' Images of Motion. *Music Perception, 23*, 221-248.

Eitan, Z., Ornoy, E., & Granot, R. Y. (2012). Listening in the dark: Congenital and early blindness and cross-domain mappings in music. *Psychomusicology: Music, Mind, and Brain*, *22*(1), 33.

Erdemir, A., Erdemir, E., & Rieser, J. (2010) A kinematic model for perceived musical tempo. *Proceedings of the International Conference on Music Perception and Cognition, August 23-27, Seattle, WA*.

Erdemir, A., & Rieser, J. J. (2016). Singing without Hearing: The Use of Auditory and Motor Information when Singers, Instrumentalists and Non-musicians Sing a Familiar Tune. *Music Perception, 33(5).*

Etchell, A. C., Johnson, B. W., & Sowman, P. F. (2014). Beta oscillations, timing, and stuttering. *Frontiers in human neuroscience*, *8*.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology, 73*, 2608–2611.

Fadiga L, Craighero L, & Olivier E. (2005). Human motor cortex excitability during the perception of others' action. *Current Opinion in Neurobiology 15*:213–218.

Falk, S., Müller, T., & Dalla Bella, S. (2015). Non-verbal sensorimotor timing deficits in children and adolescents who stutter. *Frontiers in Psychology*, *6*.

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature, 424*, 769–771.

Friberg, A., & Sundberg, J. (1999). Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America, 105*, 1469–1484.

Gallese V, Fadiga L, Fogassi L,& Rizzolatti G. (1996). Action recognition in the premotor cortex. *Brain 119*:593–609.

Gangitano M, Mottaghy FM, Pascual-Leone A. 2001. Phase specific modulation of cortical motor output during movement observation. *NeuroReport 12*:1489–92.

Gardner, J. M., Lewkowicz, D. J., Rose, S. A., & Karmel, B. Z. (1986). Effects of visual and auditory stimulation on subsequent visual preferences in neonates. *International Journal of Behavioral Development, 9*(2), 251-263.

Gibson JJ (1996) The Senses Considered as Perceptual Systems. Boston: Houghton-Mifflin.

Gogate, L. J., Bahrick, L. E. & Watson, J. D. (2000) A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development 71*:878–94.

Gordon, R. L., Jacobs, M. S., Schuele, C. M., & McAuley, J. D. (2015). Perspectives on the rhythm–grammar link and its implications for typical and atypical language development. *Annals of the New York Academy of Sciences*, *1337*(1), 16-25.

Gordon, R. L., Shivers, C. M., Wieland, E. A., Kotz, S. A., Yoder, P. J., & Devin McAuley, J. (2015). Musical rhythm discrimination explains individual differences in grammar skills in children. *Developmental science*, *18*(4), 635-644.

Grahn JA, Brett M. (2007). Rhythm and beat perception in motor areas of the brain. J *Cognitive Neuroscience. 19*:893-906.

Gramming, P., Sundberg, J., Ternstrom S., Leanderson, R., & Perkins, W. H. (1988) Relationship between changes in voice pitch and loudness. *Journal of Voice 2*, 118–126.

Harlan D. P., (n.d.) Laban Movement Analysis/Bartenieff Fundamentals.

Hasegawa, A., Okanoya, K., Hasegawa, T., & Seki, Y. (2011). Rhythmic synchronization tapping to an audio-visual metronome in budgerigars. *Scientific Reports*, 1, doi:10.1038/srep00120

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of Acoustical Society of America 97*, 3099–3111.

Hommel, B., Müsseler, Aschersleben, G. & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849-937.

Hurley, S. & Chater, N. (1995) *Perspective on Imitation: From Neuroscience to Social Science.* MIT Press, Cambridge, Massachusetts.

Iacoboni M, Woods RP, Brass M, Bekkering H, Mazziotta JC, & Rizzolatti G. (1999). Cortical mechanisms of human imitation. *Science 286*: 2526–28.

James, W. (1890). The Principles of Psychology. Dover, New York. Prinz, W. (1997). Perception and action planning. *Eur. J. Cognitive.Psychology. 9*: 129–154.

Jeannerod M. (2006) *Motor cognition: what action tells the self.* New York: Oxford University Press.

Johansson G, (1973) Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics* 14 201-211.

Iohnstone, T., & Scherer, K. (2000). Vocal communication of emotion. *Handbook of emotion*, 220-235.

Jones, J. A., & Keough, D. (2008). Auditory-motor mapping for pitch control in singers and nonsingers. *Experimental Brain Research, 190*, 269–287.

Jones, J. A., & Munhall, K. G. (2002). The role of auditory feedback during phonation: studies of Mandarin tone production. *Journal of Phonetics*, *30*(3), 303-320.

Kawahara, H. (1995) Hearing voice: transformed auditory feedback effects on voice pitch control. *Proceedings of the international joint conference on artificial intelligence: workshop on computational auditory scene analysis,* pp. 143–148. Montreal, Canada

Kilner, J. M., Paulignan, Y., & Blakemore, S.-J. (2003). An interference effect of observed biological movement on action. *Current Biology, 13*(6), 522_525.

Kohler E, Keysers C, Umilta` MA, Fogassi L, & Gallese V (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science 297*:846–848.

Kohn, D., & Eitan, Z. (2009) Musical Parameters and Children's Movement Responses. *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music*. pp.233-241.

Kohn, D., & Eitan, Z. (2012) Seeing Sound Moving: Congruence of Pitch and Loudness with Human Movement and Visual Shape. *Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and 8th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM),* p. 541.

Kozlowski L. T., & Cutting J. E., (1977) Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics 21*: 575-580.

Küssner, M. B., Tidhar, D., Prior, H. M., & Leech-Wilkinson, D. (2014). Musicians are more consistent: Gestural cross-modal mappings of pitch, loudness and tempo in real-time. *Frontiers in psychology*, *5*. 789.

Küssner, M. B., & Leech-Wilkinson, D. (2013). Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm. *Psychology of Music*, 42(3):448-469.

Laban, R, & Lawrence, F. C. (1947) *Effort*. London: Macdonald and Evans.

Lahav A, Saltzman E, & Schlaug G (2007) Action representation of sound: audiomotor recognition network while listening to newly acquired actions. *Journal of Neuroscience 27*: 308–314.

Lane, H. L., & Tranel, B. (1971) The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research, 14*, 677-709.

Lane, H., & Webster, J. W. (1991). Speech deterioration in postlingually deafened adults. *Journal of the Acoustical Society of America, 89,* 859-866.

Leder, S. B., Spitzer, J. B., & Kirchner, J. C. (1987a). Speaking fundamental frequency of postlingually profoundly deaf adult men. *Annals of Otology, Rhinology and Laryngology, 96,* 322-324.

Leder, S. B., Spitzer, J. B., Kirchner, J. C., Flevaris-Phillips, C., Milner, P., & Richardson, F. (1987b). Speaking rate of adventitiously deaf male cochlear implant candidates. *Journal of the Acoustical Society of America, 82,* 843-846.

Leder, S. B., Spitzer, J. B., Milner, P., Flevaris-Phillips, C., Kirchner, J. C., & Richardson, F. (1987c). Voice intensity of prospective cochlear implant candidates and normal hearing adult males*, Laryngoscope, 97*, 224-227.

Lewkowicz DJ, Leo I, & Simion F. (2010) Intersensory perception at birth: Newborns match non-human primate faces & voices. *Infancy, 15*:46–60.

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory-visual intensity matching. *Developmental Psychology, 16*, 597-607.

Lewkowicz, D. J., & Turkewitz, G. (1981).Intersensory interaction in newborns: Modification of visual preferences following exposure to sound. *Child Development, 52*, 827-832.

Lipscomb, S.D., & Kim, E.M. (2004). Perceived match between visual parameters and auditory correlates: An experimental multimedia investigation. *Proceedings of the 8th International Conference on Music Perception & Cognition (ICMPC8)*, pp.72-75.

Luck, G. P. B. & Toiviainen, P. (2006). Ensemble musicians' synchronization with conductors' gestures: an automated feature-extraction analysis. *Music Perception, 24*(2), 195-206.

Maeda F, Kleiner-Fisman G, & Pascual-Leone A. (2002). Motor facilitation while observing hand actions: specificity of the effect and role of observer's orientation. *Journal of Neurophysiology. 87*:1329–35.

Mather G., & Murdoch L. (1994) Gender discrimination in biological motion displays based on dynamic cues" *Proceedings of the Royal Society of London, Series B 258,* 273-279.

Martinez-Castilla, P., & Sotillo, M. (2008). Singing Abilities in Williams syndrome. *Music Perception: An Interdisciplinary Journal, 25*(5), 449–469.

Meltzoff AN, & Prinz W. (2002). *The Imitative Mind. Development, Evolution and Brain Bases*. Cambridge, UK: Cambridge Univ. Press.

Mendelson, M. J., & Ferland, M. B. (1982).Auditory-visual transfer in four-month-old infants. *Child Development, 53*, 1022-1027.

Montagna, M., Cerri, G., Borroni, P., & Baldissera, F. (2005).Excitability changes in human corticospinal projections to muscles moving hand and fingers while viewing a reaching and grasping action. *European Journal of Neuroscience 22*, 1513–1520.

Montgomery, K. J., Isenberg, N., & Haxby, J. V. (2007).Communicative hand gestures and object-directed hand movements activate the mirror neuron system. *Social, Cognitive & Affective Neuroscience, 2*, 114–122.

Morrongiello, B. A., Fenwick, K. D., & Chance, G. (1998). Crossmodal learning in newborn infants: Inferences about properties of auditory-visual events. *Infant Behavior and Development, 21*(4), 543-554.

Mürbe, D., Pabst, F., Hofmann, G., & Sundberg, J. (2002). Significance of auditory and kinesthetic feedback to singers' pitch control. *Journal of Voice*, *16*(1), 44-51.

Mürbe, D., Pabst, F., Hofmann, G., & Sundberg, J. (2004). Effects of a professional solo singer education on auditory and kinesthetic feedback – a longitudinal study of singers' pitch control. *Journal of Voice*, *18*(2), 236-241.

Nymoen, K., Godøy, R. I., Jensenius, A. R., & Torresen, J. (2013). Analyzing correspondence between sound objects and body motion. *ACM Transactions on Applied Perception (TAP)*, *10*(2), 9.

Patel, A.D. et al. (2009) Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current Biology 19*, 827–830.

Prieto, P., & Ortega-llebaria, M. (2006). Stress and Accent in Catalan and Spanish: Patterns of duration, vowel quality, overall intensity, and spectral balance, in R. Hoffmann and H. Mixdorff (eds.), *Proceedings of Speech Prosody*, 337-340.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, 9, 129-154.

Phillips-Silver, J., & Trainor, L.J. (2008). Vestibular influence on auditory metrical interpretation. *Brain and Cognition, 67*, 94-102.

Raphael, L. J, Borden, G.J., & Harris, K.S. (2007). Speech science primer: Physiology, acoustics, and perception of speech (5th ed.). Philadelphia: Lippincott Williams & Wilkins.

Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. Annual Review of *Neuroscience, 27*, 169-92.

Roffler, S. K., & Butler, R. A. (1968). Localization of tonal stimuli in the vertical plane. *Journal of the Acoustical Society of America, 43,* 1260-1265.

Schachner, A., Brady, T., Pepperberg, I., & Hauser, M. (2009). Spontaneous Motor Entrainment to Music in Multiple Vocal Mimicking Species. *Current Biology*, 19, 831–836.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, *40* (1), 227-256.

Schnupp J, Nelken I, & King AJ. (2011) *Auditory Neuroscience: Making Sense of Sound*. Cambridge, MA: MIT Press.

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., & Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. *Current opinion in neurobiology*, *20*(2), 172-176.

Schultz-Coulon, H. J. (1978). The neuromuscular phonatory control system and vocal function. *Acta Otolaryngol*, *86*, 142–153.

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception, 36*, 888-897.

Slater, A., Quinn, P. C., Brown, E., & Hayes, R. (1999). Intermodal perception at birth: Intersensory redundancy guides newborn infants' learning of arbitrary auditory-visual pairings. *Developmental Science, 3*, 333-338.

Su, Y.-H., & Jonikaitis, D. (2011). Hearing the speed: visual motion biases the perception of auditory tempo. *Experimental Brain Research, 214*, 357–371.

Sundberg, J., & Verrillo, V. (1980). On the anatomy of the ritard: A study of timing in music. *Journal of Acoustical Society of America 68*, 772–779.

Sperry, R.W. (1952). Neurology and the mind-body problem. *American Scientist, 40*, 291-312.

Thompson, M. R., & Luck., G. (2008) Exploring relationships between expressive and structural elements of music and pianists' gestures. *Proceedings of the fourth Conference on Interdisciplinary Musicology, Thessaloniki, Greece.*

Ticini LF, Schutz-Bosbach S, Weiss C, Casile A, & Waszak F (2012) When sounds become actions: higher-order representation of newly learnt action sounds in the human motor system. *J Cogn Neurosci 24(*2):464–474.

Todd, N. P. & McA. (1992) The dynamics of dynamics: A model of musical expression, *Journal of Acoustical Society of America*. *91*, 3540–3550.

Wallace, M. T., Ramachandran, R., & Stein, B. E. (2004). A revised view of sensory cortical parcellation. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(7), 2167-2172.

Wan, C. Y., Bazen, L., Baars, R., Libenson, A., Zipse, L., Zuk, J.& Schlaug, G. (2011). Auditory-motor mapping training as an intervention to facilitate speech output in non-verbal children with autism: a proof of concept study. *PloS one*, *6*(9), e25505.

Walker, P., Bremner, J.G., Mason, U., Spring, J., Mattock, K., & Slater, A. (2010).Preverbal infants' sensitivity to synesthetic cross-modality correspondences. *Psychological.Sciences 21*, 21–25.

Ward, W. D., & Burns, E. M. (1978). Singing without auditory feedback. *Journal of Research in Singing & Applied Vocal Pedagogy*, *1*, 24-44.

Wieland, E. A., McAuley, J. D., Dilley, L. C., & Chang, S. E. (2015). Evidence for a rhythm perception deficit in children who stutter. *Brain and language*,*144*, 26-34.

Widmann, A., Kujala, T., Tervaniemi, M., Kujala, A., &Schroeder, E. (2004). From symbols to sounds: Visual symbolic information activates sound representations. *Psychophysiology, 41,* 709-715.

van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration: flexible use of general operations. *Neuron*, *81*(6), 1240-1253.

Viviani, P., & Stucchi, N. (1989). The effect of movement velocity on form perception: Geometric illusions in dynamic displays. *Perception & Psychophysics, 46*, 266-274.

Viviani P, Stucchi N. (1992) Biological movements look uniform: evidence for motor-perceptual interactions. *Journal of Experimental Psychology: Human Perception & Performance; 18*: 603–23.

Zbokowski, L. (1998). Metaphor and music theory. Music Theory Online, 4. Retrieved from http://societymusictheory.org/mto/issues/mto.98.4.1/mto.98. 4.1.zbikowski.html

Zentner, M., & Eerola, T. (2010). Rhythmic engagement with music in infancy. *Proceedings of the National Academy of Sciences of the United States of America, 107*(13), 5768–5773.