Relations between music and speech from the perspectives of dynamics, timbre and pitch

Xiaoluan Liu

A thesis submitted in fulfilment of requirements for the degree of

Doctor of Philosophy

to

Department of Speech, Hearing and Phonetic Sciences

Division of Psychology and Language Sciences

University College London (UCL)

2016

Declaration

I, Xiaoluan Liu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. Portions of this thesis have been or are likely to appear published in peer-reviewed journals and conference proceedings.

Acknowledgements

My deepest gratitude goes to Professor Yi Xu, my primary supervisor, for his enthusiastic support for my project. His intelligence, insight, critical thinking and broad knowledge have always been a great source of inspiration to me. His encouragement for me to always think broadly and deeply has led to the publication of some studies of this thesis. Without his guidance and support, this would have been impossible. I've also benefited a great deal from the intellectual guidance from Dr Jyrki Tuomainen, who introduced me to the world of neuropsychology, especially the ERP technique. His patience and great attention to detail have helped me gain a better understanding of ERP in the context of music and speech science. My second supervisor, Dr Bronwen Evans, is always there to give great support and encouragement for me, especially at the time when I seriously doubted the prospect of my research. My sincere gratitude also goes to Dr Andrew McPherson at Queen Mary University of London, for his generous help with piano data acquisition using PianoBar. Without his support, the first study and the publication of the study would have been impossible.

Meanwhile, I am also indebted to Professor Paul Iverson, Professor Mark Huckvale, Professor William Forde Thompson, and Dr Kai Alter for their interest in my research. Some of the ideas of this thesis were the results of informal discussions with them. I would also like to thank my fellow PhD friends, Tim Schoof, Emma Brint, Louise Stringer and Kurt Steinmetzger for their great help with some of the technical aspects of ERP recording. I also enjoyed the discussions and great time spent together with other PhD friends (Albert Lee, Chris Hsu, Danniel Kennedy-Higgins, Faith Chiu, Gise Lourido, Hao Liu, Jieun Sung, Mauricio Candia, Yue Zhang). Also, I can never forget the help from the participants for my research.

Without their patience, understanding, and willingness to sacrifice their valuable time, this thesis would have never been finished.

My family, my grandma in particular, have always been my strongest supporters who have never failed to give me moral and intellectual guidance since I was born. Their intelligence, humour, and resilience will always inspire me to achieve my goals in life. Therefore, this thesis is dedicated to them, especially my grandma.

Abstract

Despite the vast amount of scholarly effort to compare music and speech from a wide range of perspectives, some of the most fundamental aspects of music and speech still remain unexplored. This PhD thesis tackles three aspects essential to the understanding of the relations between music and speech: dynamics, timbre and pitch. In terms of dynamics, previous research has used perception experiments where dynamics is represented by acoustic intensity, with little attention to the fact that dynamics is an important mechanism of motor movements in both music performance and speech production. Therefore, the first study of this thesis compared the dynamics of music and speech using production experiments with a focus on motor movements: finger force in affective piano performance was used as an index of music dynamics and articulatory effort in affective Mandarin speech was used as an index of speech dynamics. The results showed both similarities and differences between the two domains. With regard to timbre, there has been a long-held observation that the timbre of musical instruments mimics human voice, particularly in terms of conveying emotions. However, little research has been done to empirically investigate the emotional connotations of the timbre of isolated sounds of musical instruments in relation to affective human speech. Hence, the second study explored this issue using behavioral and ERP methods. The results largely supported previous observations, although some fundamental differences also existed. In terms of pitch, some studies have mentioned that music could have close relations with speech with regard to pitch prominence and expectation patterns. Nevertheless, the functional differences of pitch in music and speech could also imply that speech does not necessarily follow the same pitch patterns as music in conveying prominence and expectation. So far there is little empirical evidence to

either support or refute the aforementioned observations. Hence the third study examined this issue. The results showed the differences outweighed the similarities between music and speech in terms of pitch prominence and expectation. In conclusion, from three perspectives essential to music and speech, this thesis has shed new light on the overlapping yet distinct relations between the two domains.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Music and speech reflect fundamental aspects of human capacities (Patel, 2008). The relations between music and speech have been attracting scholarly interest for a long time (Fonagy and Magdics, 1963; Scherer, 1995; Sundberg, 1982), with attempts to compare the two from a wide range of perspectives: prosody (Scherer, 1995), semantics (Seifert et al., 2013), syntax (Lerdahl, 2013), evolution (Cross et al., 2013), neurocognitive mechanisms (Steinbeis and Koelsch, 2008) and facial expressions (Carlo and Guaitella, 2004; Livingstone et al., 2015).

Particularly, an increasing amount of attention has been given to comparisons between affective music and speech. The reason is that vocal expression of emotion is a crucial aspect of speech communication, which has been found to exist cross-culturally (Scherer, 2003); communicating emotion is also a primary function of music that exists in both music performance and music perception (Juslin and Laukka, 2003). Therefore, music and speech have been two important channels for humans to communicate emotion (Juslin and Laukka, 2003). Early attempts to find parallels between the two domains could be found in Spencer (1857) where singing was argued to be closely associated with vocal expression of emotions, according to the hypothesis that emotion triggers physiological changes which could be the precondition for the acoustic correlations between music and speech. This classic argument has served as the stepping stone for subsequent scholarly endeavor for the

search of the similarities and differences between affective music and speech (Scherer, 1995; Sundberg, 1982).

The majority of studies on the comparisons between the two domains show that perceptually, acoustic cues (pitch, intensity and duration) of affective music and speech are similar (Curtis and Bharucha, 2010; Ilie and Thompson, 2006; Juslin and Laukka, 2003). For example, in highly activated emotions such as anger and happiness, the sound intensity increases and the high-low frequency energy ratio increases; the opposite is true in much less activated emotions such as sadness and tenderness (Juslin and Laukka, 2003). Acoustic differences between the two domains also exist, but on the whole similarities outweigh differences according to the seminal review of around 150 studies on affective music and speech (Juslin and Laukka, 2003).

1.2 A brief introduction to the aims of this thesis

Past research as outlined above has brought valuable insight into the relations between music and speech. However, some fundamental aspects of music and speech still remain insufficiently explored especially in terms of dynamics, timbre and pitch. With regard to dynamics, research is sparse on the comparisons between the dynamics of music and speech from the perspective of production paradigms. This could be due to the fact that compared to the vast amount of perception studies, research using production paradigms on music is on the rise only in the past twenty years, thanks to the advent of music technology that makes quantifying music production (i.e., performance) easier than before. Particular interest has been given to affective piano/keyboard performance due to the availability of MIDI, 3D motion capture cameras and digital acoustic pianos (Palmer, 2006). Nevertheless, to our

knowledge, strictly controlled experiments that directly compare the dynamics of affective piano performance with those of affective speech production are rare. Given the fact that both speech production and piano performance are essentially human motor movements, it is worth investigating if the two domains are similar or different with regard to the mechanisms of motor movements. Dynamics is an important mechanism of motor movements (more details and justification of selecting dynamics as a measurement parameter will be given in Chapter 2) which so far has not been systematically examined in the comparisons between music performance and speech production. Therefore, the first study of this thesis aimed to compare music and speech from the perspective of affective piano performance and speech production with a special focus on dynamics.

The second aspect that is not well explored in the relations between music and speech is timbre. Timbre is a crucial acoustic dimension of sounds and its importance in conveying emotion in both music and speech is undeniable: in music, timbre has been an effective platform enabling composers to induce emotions from listeners (Boulez, 1987; Gabrielsson, 2001) because of its significant role in the expression and perception of emotion (Eerola et al., 2012; Patel, 2008). In speech, timbre (i.e., voice quality) is an important means of vocal communication of affect, especially in terms of conveying fine-tuned affective states such as confidence, boredom, formality, etc. (Gobl and Ní Chasaide, 2003). Despite the importance of timbre, it is not as sufficiently studied as other acoustic features such as pitch, duration and intensity (Patel, 2008). This is especially true with regard to exploring the emotional connotations of timbre of isolated musical instrument sounds (Eerola et al., 2012). Thus it is worth further exploring this issue, particularly with affective speech as a reference. This is because musical instruments have long been compared

to human voice due to their timbral similarities: one of the most well-known analogies is the comparison of string instruments such as the violin or the guitar to human voice due to their similar expressiveness (Askenfelt, 1991). Nevertheless, so far there have been no strictly controlled behavioral or neurophysiological studies on the relations between the timbre of isolated musical instrument sounds and affective human voice. Hence, the second study of this thesis explored this topic with behavioral and neurophysiological (ERP) methods (more reviews and justifications will be provided in Chapter 3).

The third aspect worthy of further examination is pitch prominence and expectation in music and speech. Pitch prominence plays an important role in both music and speech. Prosodically prominent elements such as focus in speech or melodic accent in music can direct listeners' attention to the emphasized elements of acoustic signals (either of speech or music) to facilitate comprehension. This is because the ability to perceive and interpret abstract cues such as focus or accent reflects one of the fundamental aspects of comprehension, i.e., differentiation between the important and the unimportant (Balkwill and Thompson, 1999). Therefore, pitch prominence could be one of the crucial aspects of acoustic communication in music and speech (Parncutt, 2003), both of which boil down to attracting greater perceptual weight ('t Hart et al., 1990) than non-prominent prosodic elements (Benward and White, 1997). Another aspect of pitch variation employed in both speech and music is related to expectation. Expectation is part of psychological laws of mental life responsible for human perception and cognition (Meyer, 1956). More specifically, it is a cognitive mechanism enabling humans to make predictions about the development of future events (Meyer, 1956). Violation of expectation can trigger the emotion of surprise, which can be found in both speech and music, i.e., both domains use pitch variation

to convey surprise (Meyer, 1956). As informally suggested in many studies (e.g., Huron, 2006; Parncutt, 2003; Patel, 2008), speech could follow the same pitch patterns in conveying prominence and expectation as those in music. Nevertheless, the functional differences in the use of pitch between music and speech may suggest otherwise (Peretz and Hyde, 2003). It is thus worth further examining this issue by testing whether under laboratory conditions, music and speech show similar or different pitch patterns with regard to prominence and expectation. So far no research along this line has been done. Therefore, the third study of this thesis addressed this issue by manipulating the pitch patterns of speech and music in terms of prominence and expectation (more details and motivations for this study will be elaborated on in Chapter 4).

1.3 Bio-evolutionary implications of music and speech

Another aim of the thesis is to explore the bio-evolutionary implications of music and speech through the lenses of dynamics, timbre and pitch as introduced above. The first platform for examining such implications is emotion. This is because music and speech are important channels for humans to communicate emotion (Juslin and Laukka, 2003). Emotion, from an evolutionary perspective, is adapted under selection pressure (Darwin, 1872). As Ekman (1992) once commented, emotion is one of the mechanisms for living organisms to interact with one another: "the primary function of emotion is to mobilise the organism to deal quickly with important interpersonal encounters, prepared to do so in part, at least, by what types of activity have been adaptive in the past" (p.171).

Correspondingly, emotional vocal expressions are likely selected to have the effect of influencing the receiver for the benefit of the signaller (Morton, 1977; Ohala,

1984; Xu et al. 2013a, 2013b). One particular line of research has taken such a bio-evolutionary perspective in studying emotional speech, based on the body size projection theory on emotion. The theory was originally proposed by Morton (1977) for explaining animal calls, and later extended by Ohala (1984) to human speech. The key idea is that vocal emotional expressions are a mechanism evolved under a selection pressure to influence the behaviour of other individuals in social interactions. For example, (hot) angry vocalizations signal a large body size projection because evolutionarily a large animal stands a better chance of winning physical confrontations (i.e., the "fight" response triggered by anger), while happy (joyful) vocalizations signal a small body size projection because evolutionarily a small animal or an infant could suggest attractiveness and lack of threat (Morton, 1977; Ohala, 1984). Recently, this idea has seen support from a series of perception experiments in which the speech stimuli are synthetically manipulated in terms of pitch, vocal tract length and voice quality to simulate body size projection (Chuenwattanapranithi et al., 2008; Xu et al., 2013a, 2013b). It is shown that listeners hear speech with synthetic parameters that project a large body size both as being spoken by a large person and as expressing anger. And they hear speech that projects a small body size as spoken by a small person and expressing happiness and friendliness (Noble and Xu, 2011; Xu et al., 2013a, 2013b).

Of particular relevance to the first study of this thesis is the "dynamics" dimension of the bio-evolutionary account on emotion (Xu et al., 2013a). Dynamics, in the context of speech, reflects the extent of effort/vigour of human vocalizations (Xu et al., 2013a): high dynamics corresponds to a high extent of vocal effort while a low dynamics corresponds to a low extent of vocal effort. (Hot) anger and happiness (joy) are predicted to have high dynamics (Xu et al., 2013a). This is because anger is a

high-arousal emotion associated with a high level of bio-physical activation which could help demonstrate great energy and strength to scare off enemies (Morton, 1977; Xu et al., 2013a, 2013b). As a result, the extent of vocal effort (i.e., dynamics) should be high in anger. Happiness also links to high dynamics because from an evolutionary perspective, happiness can be a useful strategy for attracting mates (Darwin, 1872). Therefore, it is beneficial for sound signallers to produce highly vigorous (i.e., dynamic) sounds so as to be audible to potential mates (Xu et al., 2013a). Correspondingly, vocal effort in happiness should correlate to high dynamics. With regard to fear, it should also have reasonably high dynamics due to the evolutionary function of fear: it could be an antipredator defensive strategy for group survival used by many species (Caro, 2005; LeDoux, 1996). Therefore, vocal effort should be reasonably high to serve this purpose. In terms of (depressed) sadness, the dynamics level should be low due to the low level of arousal and bio-physical activation. Consequently, it could send out an evolutionary signal of begging for sympathy (Shaver et al., 1987). As introduced in the previous section, dynamics is an essential mechanism of motor movements (Stein, 1982). Since both speech production and piano performance are motor movements, plus the fact that both of them are primary platforms for humans to communicate emotion, it is therefore reasonable to test (in Chapter 2) whether emotional speech production and piano performance follow similar dynamics patterns (reflected by vocal effort and finger force respectively) according to the aforementioned bio-evolutionary predictions.

The second study of this thesis is focused on timbre. According to the hypotheses of body size projection theory on emotional vocalizations, (hot) angry vocalizations should correspond to rough timbre reflected by an abundance of high frequency

spectral energy (Morton, 1977; Xu et al., 2013a, 2013b). This is because anger, according to the theory, projects a large body size in order to scare off enemies in physical confrontations. Due to simple physical laws, the vocalization of a large-size animal is likely to have rough sound quality (Morton, 1977). Happiness (joy), on the other hand, projects a small body size with evolutionary signals of attracting mates and showing a lack of threat (Morton, 1977). As a result, the timbre of happy vocalizations often sounds like a pure tone characterized by reduced high frequency spectral energy (Xu et al., 2013a, 2013b). (Depressed) sadness could project a neutral or small body size due to its evolutionary signal of begging for sympathy (Shaver et al., 1987). Therefore, the timbre of sad vocalizations should be associated with reduced high frequency spectral energy due to the low physical activation level in articulation (Xu et al., 2013a). As introduced in the previous section, musical instrument timbre can be compared to human voice quality, and both music and speech could have evolutionary implications through the lens of emotion (Darwin, 1871; Cross, 2009). It is therefore worth examining whether musical instrument timbre has acoustic characteristics that convey emotion in the same direction as predicted by the body-size projection theory for emotional vocalizations. This will be tested by the behavioural experiment in Chapter 3. The following two ERP experiments will demonstrate whether brain responses to the musical instrument timbre will show similar patterns to those of affective speech, thus further testing whether musical instrument timbre could communicate emotion in a way similar to affective speech. This could further imply whether the brain processing patterns could be consistent with the predictions of the body size projection theory on emotional vocalizations. This is because the brain responses are triggered by participants' evaluation of the timbral features of the musical sounds, and if the

timbral features are consistent with predictions of the body size projection theory (as will be tested by the behavioural experiment), then the brain responses triggered by the timbre will also reflect (indirectly) the brain's evaluation of the sounds in the direction as predicted by the body size projection theory.

Emotion is not the only platform to investigate the bio-evolutionary implications of music and speech. Codified meaning in speech and music is another dimension to examine such implications. Intonation, as proposed in Gussenhoven (2004), can be explained from the perspective of "biological codes" (i.e., the effort code, frequency code, and production code). Such a biological perspective on the codified meaning in speech has also been applied to analyse the relationship between language and music by Cross and Woodruff (2009). Of particular relevance to this thesis is the "effort code" (Gussenhoven, 2004), which could be applied to explain pitch prominence (the topic of Chapter 4). The effort code is based on the idea that effortful articulation is associated with less slurring and hence less target undershoot than sloppy articulation. This usually results in more precise articulation and a wider pitch expansion (Gussenhoven, 2004). Therefore, expanded pitch range can be used to signal informational prominence (e.g., emphasis, focus) (Gussenhoven, 2004). The mechanism of the effort code in speech can also be applied to music melodic contours, because "(melodic) peaked contours might serve to highlight ostensively certain features of a musical utterance, a function analogous to that of focus in speech prosody" (Cross and Woodruff, 2009: 91). Therefore, it is worth further exploring the aforementioned observations with empirical experiments, with the aim to test whether speech and music follow similar pitch patterns in signalling prominence in the same direction as predicted by the effort code. This can serve as a further test to show whether codified meanings in speech and music could also

convey bio-evolutionary implications, since the effort code is derived from the biological mechanisms of speech production.

1.4 Outline of this thesis

This thesis addresses the relations between music and speech by focusing on the three aspects (dynamics, timbre and pitch) that are not well explored in previous studies. Chapter 2 focuses on dynamics in the comparison between affective piano performance and speech production. Chapter 3 addresses the emotional connotations of musical timbre of isolated instrument sounds through the perspective of affective speech. Chapter 4 is concerned with pitch prominence and expectation in music and speech. Chapter 5 concludes this thesis with a summary of the results, discussion of the implications and suggestions for future research.

Chapter 2

Relations between affective music and speech: Evidence from dynamics of affective piano performance and speech production

2.1 Introduction

As mentioned in Chapter 1, this chapter is concerned with comparing affective music and speech from a different perspective by using affective piano performance and speech production with a special focus on dynamics. The results will be discussed in terms of the bio-evolutionary account on emotion as introduced in Chapter 1. In addition, physical constraints such as fingerings in piano performance and articulatory constraints (e.g., articulatory pressure and distance) in speech production will also be included as a factor. The reasons will be provided in the following sections.

2.1.1 Dynamics of piano performance

In studies of affective piano performance, dynamics have received less attention than timing, although they are equally important (Repp, 1996; Gabrielsson, 2003). The reason is that unlike timing which can be easily measured by metronome and hence has been systematically examined in a scientific way for over a decade (Repp, 1992a, 1992b, 1994a, 1994b, 1995, among others), dynamics are more difficult to measure. This could be partly due to perceptual difficulty in precisely distinguishing different levels of dynamics (e.g., *forte* and *mezzoforte*) or technical challenge in filtering out unwanted acoustic artefacts (Repp, 1996).

Therefore, in this study we decided to examine piano dynamics from a different perspective, i.e., the kinematic level of dynamics which reflects "the varying forces

of the pianist's finger movements on the keyboard" (Repp, 1996, p. 642) by using a modified Moog PianoBar scanner (cf. McPherson, 2013). It is a portable scanner that can be rapidly attached to any acoustic piano keyboards. Using optical reflectance sensing mechanism, the modified PianoBar scanner continuously detects key movements. Quantitatively, the scanner returns the values of continuous key positions (key displacement) and the time taken for fingers to reach each key position during one keystroke. As a result, multiple different dimensions of each key press, velocity and peak velocity (i.e., the maximum value in a continuous velocity trajectory) of key movement during each keystroke can be extracted from continuous key position data, following a similar approach to McPherson and Kim (2011). The multidimensions of key touch quantitatively returned by the scanner can provide an ideal platform for examining the interaction between pianists' expressive intention and their piano key touch/finger force (cf. McPherson and Kim, 2013).

Literature on mechanics of skilled motor movement (such as speech production and music performance) suggests that dynamics of motor movement are related not only to peak velocity but also to the movement amplitude, i.e., the peak velocity should be divided by the movement amplitude in order to compare dynamics of movement of different sizes (Nelson, 1983; Ostry et al., 1983; Ostry and Munhall, 1985). Therefore, in the context of piano performance, since each keystroke may correspond to different degrees of key displacement (i.e., different amplitudes of key movement), it is necessary to factor in key displacement at the point of peak velocity to yield the kinematic dynamics of each keystroke which reflects pianists' finger force (Minetti et al., 2007). Similar approach can also be found in Kinoshita et al. (2007) where key displacement was taken as a factor in comparing finger force under the conditions of different types of key touch.

The examination of kinematic dynamics of pianists' finger force needs to take into account the role of fingerings. This is because in piano performance different fingering strategies can reflect how pianists intend to interpret the structure, meaning and emotion of music in which dynamics play an important role (Bamberger, 1976; Clarke et al., 1997; Neuhaus, 1973). Parncutt et al. (1997) established a set of hypothetical rules of right-hand fingerings according to ergonomic difficulty such as the extent of hand spans, the involvement of weak fingers, and the coordinated playing on black and white keys. Of particular importance are hand spans and finger strength (i.e., the contrast between weak and strong fingers). This is because the extent of hand spans can affect the degree of tension and physical effort of fingers (Parncutt et al., 1997). Weak fingers usually refer to the fourth and fifth fingers (Parncutt et al., 1997) which can constrain the flexibility of finger movement because of the hand's anatomical structure: unlike the other fingers which can move relatively independently, the ring and little finger are closely linked to each other via the flexor digitorum profundus (FDP) tendons because they share a common muscle belly (Gunter, 1960). Moreover, the flexor digitorum superficialis (FDS) is especially responsible for the coupling between the fourth and fifth fingers (Baker et al., 1981; Austin et al., 1989). Nevertheless, whether weak fingers can significantly influence piano performance is still a matter of debate. As pointed out in Kochevitsky (1967), Neuhaus (1973) and Sandor (1981), weak fingers are not necessarily weak; instead, they are often strong enough to meet the demand of different levels of playing, especially octave playing.

2.1.2 Dynamics of speech production

With regard to speech, articulatory effort which reflects "force of articulation" (Malécot, 1955) is the counterpart of finger force in piano performance. Articulatory effort is essentially a neuromuscular phenomenon. Electrochemical reaction of nerve impulses triggers the activation of articulator muscles (Kirchner, 1998). Full contraction of articulator muscles occurs when agonist muscle activity outweighs the antagonist muscle activity under the condition of repeated neuron firing (Clark and Yallop, 1990). Articulatory effort is therefore the sum action of the neuron firing of each articulator muscle (Kirchner, 1998). However, direct measurements of the neuron firing of each articulator muscle are clearly too intrusive and complex to perform. Therefore, indirect measurements have been put forth through examining phenomena related to articulatory gestures: clear speech (Uchanski, 2008), fricative closures (Lavoie, 2001), trill articulation (Padgett, 2009), assimilation (Lindblom, 1983), all of which require great articulatory effort. Correspondingly, speech production models as in Lindblom and Sundberg (1971), Westbury and Keating (1986), Kirchner (1998), have been established in an attempt to quantify articulatory effort. However, the aforementioned measurements of articulatory gestures run the risk of not capturing articulatory phenomena large enough for statistically significant differences (Kaplan, 2010); in addition, the proposed models would oversimplify the reality of speech articulation which often involves much finer details than what the models can accommodate (Kaplan, 2010).

Hence, different alternatives are worth exploring. One such example is to use formant dynamics (i.e., trajectories and velocity) as an indicator of articulatory effort (cf. Cheng and Xu, 2013). Admittedly, one could argue formant dynamics may not

be a reliable indicator of articulatory effort given the fact that there does not exist a one-to-one mapping between acoustics and articulation. Nevertheless, measurements on articulators as has been mentioned above do not capture the whole picture of articulatory movement either (cf. Cheng and Xu, 2013 for more examples and discussions). Acoustic signals, on the other hand, have been argued to provide reliable information for phonetic characteristics of segments and suprasegments with theoretical (Lindblom, 1990) and experimental evidence (Perkell et al., 2002). In addition, acoustic and articulatory measurements can produce similar dynamic patterns: the evidence is that the linear relations between F_0/formant velocity and F_0/formant movement amplitude (Xu and Wang, 2009; Cheng and Xu, 2013) in acoustics are similar to those in articulation (Kelso et al., 1985). Therefore, it is justifiable to use acoustic characteristics of formant dynamics to analyze articulatory effort.

In speech, formant patterns tend to be affected by articulatory constraints (e.g., articulatory pressure and distance) in different suprasegmental and segmental contexts (Erickson et al., 2004; Kong and Zeng, 2006). Tone languages such as Mandarin can be a typical platform for investigating articulatory pressure in different suprasegmental contexts: In Mandarin, there are five types of tones—High (tone 1), Rising (tone 2), Low (tone 3), Falling (tone 4), and Neutral (tone 5). Tone 2 + tone 2 and tone 4 + tone 4 create high articulatory pressure while tone 3 + tone 1 create low articulatory pressure. The reason is that as reported in Xu and Wang (2009), successive rising tones (i.e., tone 2 + tone 2) or falling tones (tone 4 + tone 4) create much larger articulatory pressure than other tonal combinations because they move in directly opposite tonal directions, i.e., the first tone 2 (or tone 4) ends in the opposite direction of the start of the second tone 2 (or tone 4). Successive static tones

(tone 3 and tone 1), in contrast, have much smaller articulatory pressure because the tonal movement at the end of tone 3 is in the same direction as that at the beginning of tone 1. With regard to the segmental dimension, diphthongs (i.e., two adjacent vowels) can be used because they are categorized into wide and narrow diphthongs according to their articulatory distance: wide diphthongs (e.g., /ai/, /au/, /ɔi/) have wider articulatory distance between the initial and final vowel and hence reflect greater articulatory movement of speech organs. Narrow diphthongs (e.g., /ei/, /əu/) have narrower articulatory distance between the initial and final vowel and hence the articulatory movement is not as large as wide diphthongs.

2.1.3 Motivations for this study

Theoretically, motion for a long time has been an important platform for investigating music and speech, i.e., how physical motion is associated with sound patterns subsequently generated (Sundberg, 2000). Human voice is a direct reflection of such motion-to-sound mapping through physical coordination of articulatory gestures; meanwhile, performance of musical instruments is another way of mapping motion to sound through the use of tonguing, breathing, and fingering (Palmer et al., 2007, 2009). Therefore, similar to speech production, music performance can be conceptualized as a "sequence of articulatory movements resulting in a continuous acoustic wave" (Palmer et al., 2007, p. 119). In the context of piano performance, fingers can thus be perceived as "articulators" for pianists to articulate their interpretation of music. Indeed, experimental results on piano performance (Winges et al., 2013) show that speech production phenomenon such as coarticulation also exists in pianists' finger movement during performance. This is not surprising given the fact that both speech production and piano performance are under neuromuscular

control (Winges et al., 2013) and essentially both domains require skilled motor movements following similar physical mechanisms of dynamics (Grillner et al., 1982; Nelson, 1983; Ostry, Keller and Parush, 1983; Winges et al., 2013; van Vugt et al., 2014). In the context of motor movement, force is a crucial component contributing to the dynamics of physical movements (Stein, 1982). Moreover, it is not rare to compare articulatory effort in speech with force of other types of motor movements (e.g., limb movements) (Gentil and Tournier, 1998; Ito et al., 2004; Loucks et al., 2010). As discussed in sections 2.1.1 and 2.1.2, the kinematic dynamics of keystroke reflect pianists' finger force and the formant dynamics of speech reflect speakers' articulatory effort. Since music performance and speech are two important platforms for humans to communicate emotion (Juslin and Laukka, 2003), plus the fact that these two domains are essentially skilled motor movements following similar physical mechanisms of dynamics as discussed above, it is therefore justifiable to compare music performance and speech production in the context of emotion using dynamics of motion (i.e., kinematic dynamics of keystroke and formant dynamics of speech production) as a measurement parameter. To our knowledge, such comparison is currently missing in literature and we believe it is worth bridging the gap.

In addition, one may wonder how piano fingerings (section 2.1.1) and articulatory constraints (section 2.1.2) can relate to each other. Anatomically, articulation refers to motor movements caused by skeletal muscle contraction (Tortora, 2002). Hence typical human motor movements such as speech production or music performance are effectively muscular articulation. There is no wonder, therefore, that pianists' fingers are always referred to as "articulators" expressing pianists' interpretation of music. Different fingerings involve different degrees of hand span and alternation

between strong and weak fingers, which consequently lead to different degrees of finger muscular tension (Parncutt et al., 1997). Similarly in speech, different articulatory constraints (articulatory pressure and distance) are involved as discussed in section 2.1.2. Both finger muscular tension and speech articulatory pressure/distance can be considered as physical constraints on motor movements such as piano performance and speech production (Nelson, 1983; Winges et al., 2013). Therefore, it is the physical constraints triggered by different fingerings or articulatory constraints that relate the two domains to each other. Despite the importance of fingerings and articulatory constraints reviewed in sections 2.1.1 and 2.1.2, it is still unknown whether they interact with emotion in piano performance and speech production. Meanwhile, investigating interaction between different cues is an important aspect in emotion research (cf. Juslin and Timmers, 2010). Hence, this serves as another motivation for this study.

Four of the basic emotions (Ekman, 1992) are chosen: anger, fear, happiness and sadness. One may wonder why a discrete model of emotion (Ekman, 1992; Panksepp, 1998) has been chosen rather than a dimensional approach such as Russell's circumplex model (1980). This is because firstly, so far no theoretical consensus has been reached as to which approach is better than the other for modelling emotion (for a recent summary of theoretical debates, see Zachar and Ellis, 2012). More importantly, the two approaches are not necessarily in conflict with each other as recent affective neuroscience studies (e.g., Panksepp and Watt, 2011) have suggested that the differences between the two may well be insignificant given the fact that both approaches share many common grounds in explaining cognitive functions of emotion. Since it is not the purpose of this study to test which model is better, a discrete model of affect is adopted. Among the "big six" emotions (Ekman,

1992), vocal disgust usually cannot be elicited satisfactorily under laboratory conditions (cf. Scherer, 2003); musical surprises can be very complicated often requiring sharp contrast in compositional structure (Huron, 2006) which is out of the scope of this study. Hence, only the remaining four of the "big six" emotions are chosen. The research questions to be addressed are:

Are dynamics of piano performance (i.e., finger force) similar to or different from dynamics of speech production (i.e., articulatory effort) under the condition of the four emotions? Do fingerings and articulatory constraints interact with emotion in their influence on the dynamics of piano performance and speech production respectively?

2.2 Experiment 1: The piano experiment

2.2.1 Methods

Stimuli

Two excerpts of music were composed for this study. According to the above review on fingerings, hand span and finger strength should be the primary focus. Therefore, the two excerpts were composed corresponding to two different hand spans (small vs. large) and finger strength (weak vs. strong). Small hand span is where fingers are at their natural resting positions on the keyboard, i.e., without needing to extend far beyond the resting positions to reach the notes (Sandor, 1981). Large hand span is where fingers need to extend far beyond their resting positions, which usually involves stretching at least an octave (Parncutt et al., 1997). Meanwhile, each excerpt was to be played with strong finger combinations (the thumb, index and middle fingers) and weak finger combinations (the ring and little fingers). In addition, given the fact that right and left hands tend to have different patterns in piano performance

(Minetti et al., 2007), only the right hand is involved in this experiment to avoid theoretical and practical complexities. Hence altogether there are four levels of fingerings for this study: small-weak (SW), small-strong (SS), large-weak (LW), large-strong (LS). To avoid confounding effects, all excerpts have musically neutral structure, i.e., without having overtly emotional implications. Figures 2.1-2.4 demonstrate the fingering design:

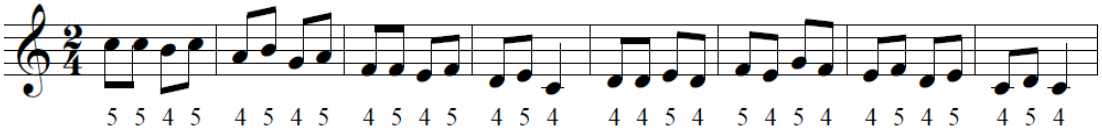


Figure 2.1 The small-weak condition (SW): small hand span (i.e., fingers are at their natural resting positions) with only weak fingers, i.e., the ring (4) and little (5) fingers involved.

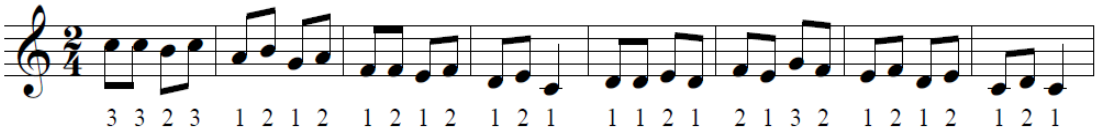


Figure 2.2 The small-strong condition (SS): small hand span (i.e., fingers are at their natural resting positions) with only strong fingers, i.e., the thumb (1), index (2) and middle (3) fingers involved.

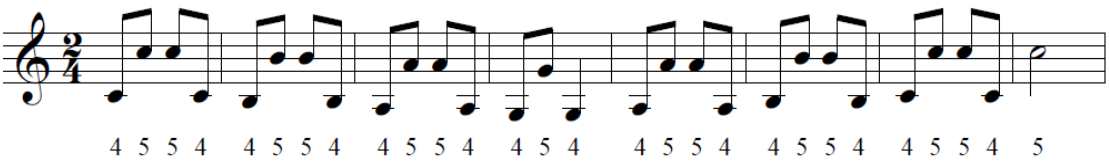


Figure 2.3 The large-weak condition (LW): large hand span (i.e., fingers stretching across an octave) with only weak fingers, i.e., the ring (4) and little (5) fingers involved.

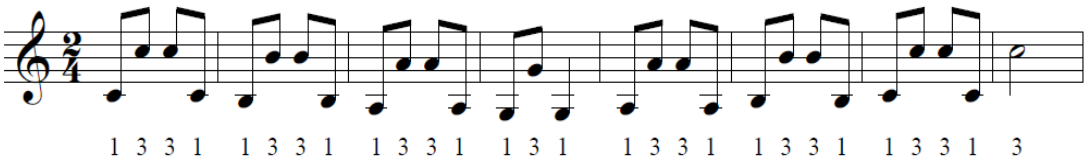


Figure 2.4 The large-strong condition (LS): large hand span (i.e., fingers stretching across an octave) with only strong fingers, i.e., the thumb (1) and middle (3) fingers involved.

Participants and procedure

This experiment was approved by the Committee on Research Ethics at University College London. Eight professional pianists (four females, Mean=26 years, SD=2.2, all right-handed) from London were recruited to play the excerpts according to the fingerings provided on scores. They have been receiving professional piano training for an average of 20 years. They were instructed to play each of the excerpts with four emotions: anger, happiness, fear and sadness. Each excerpt per emotion was repeatedly played three times in a quiet room. Admittedly, lacking ecological validity can be a problem with this method, i.e., it deviates from the reality of music making in that firstly, performances usually take place in concert halls; secondly, different emotions are often expressed by different pieces of music (cf. Juslin, 2001 for references therein). Nevertheless, real music making settings often cannot be scientifically controlled, i.e., it is impossible to filter out confounding factors coming from the acoustics of concert halls and audience. Moreover, it is hard to judge whether it is the way music is performed or the melody of music that leads the listeners to decide on the emotional categories if different pieces of music are used for different emotions (Juslin, 2000, 2001). Therefore, conducting the experiment in

a scientifically controlled way is still the better option if validity of the results is the priority.

As introduced in section 2.1.1, a Moog PianoBar scanner was attached to the keyboard of a Bösendorfer grand piano. Finger force is reflected by keystroke dynamics which were calculated according to the formula: *dynamics* = $\frac{peak\ velocity\ of\ each\ keystroke\ (Vp)}{maximum\ \ piano\ key\ displacement\ \ (d)}$ (Vp/d henceforth) because of the need to consider movement amplitude (i.e., displacement) in relation to peak velocity to reflect kinematic dynamics as reviewed in section 2.1.1. More specifically, Vp= $\frac{maximum\ piano\ key\ displacement\ (d)}{time\ taken\ to\ reach\ the\ maximum\ displacement\ (t)}$, and so Vp/d = $\frac{d}{t} \times \frac{1}{d} = \frac{1}{t}$. The unit of displacement is mm and that of time is sec. The data were obtained by an external computer attached to one end of the PianoBar. A Matlab script was written for computing dynamics according to the formula.

There were altogether 8 (pianists) * 4 (emotions) * 4 (fingerings) * 3 (repetitions) = 384 episodes performed by the pianists. A follow-up perceptual validation test was carried out: sixteen professional musicians (10 females, Mean = 28 years, SD = 1.5) were asked to rate each emotion*fingering episode on a 1 to 5 scale. 1 represented not at all angry/fearful/happy/sad while 5 represented very angry/fearful/happy/sad. The top 8 ranked episodes (out of 24) for each emotion*fingering were selected. The mean score for each emotion*fingering was 4.03.

2.2.2 Results

A two-way repeated measures ANOVA was performed to examine the effect of emotion (four levels: anger, fear, happiness and sadness) and fingerings (four levels:

small-weak, small-strong, large-weak, large-strong). The results (Table 2.1) demonstrated that both factors played significant roles in finger force reflected by Vp/d. The interaction between the two factors was also significant.

The means of keystroke dynamics (Vp/d) for each condition are displayed in Table 2.2 and Post-hoc Tukey HSD tests (Table 2.3) revealed more detailed patterns: anger and happiness had significantly higher dynamics than fear and sadness. The differences between anger and happiness were non-significant. Fear had significantly lower dynamics than anger and happiness but it was still significantly higher in dynamics than sadness. With regard to the factor of fingerings, the Tukey tests demonstrated that weak fingers in large hand span (the LW condition) did not produce significantly different dynamics from strong fingers in large hand span (the LS condition). However, under the condition of small hand span, weak fingers produced significantly lower dynamics than strong fingers.

In terms of the interaction between emotion and fingerings, Figure 2.5 shows that the most obvious interaction is between fear and different fingerings: under the conditions of large-strong (LS), large-weak (LW), and small-strong (SS) fingerings, fear had significantly higher ($p < 0.05$) dynamics than fear under the condition of small-weak (SW) fingering. Among the LS, LW and SS conditions in fear, the differences were non-significant. In addition, fear had significantly higher ($p < 0.05$) dynamics than sadness in large-strong (LS), large-weak (LW), and small-strong (SS) fingering conditions, although it was still significantly lower ($p < 0.05$) than anger and happiness. For anger, happiness and sadness, differences between fingering conditions were non-significant. This means regardless of whether the hand span was

large or small, or whether the fingers were weak or strong, the dynamics were on average always high for anger and happiness while for sadness they were always low. Therefore, the contrast in dynamics between different fingerings is evident under the condition of fear only.

Table 2.1 Mean Vp/d of the four levels of emotion and the four levels of fingerings (SW=small-weak, SS=small-strong, LW=large-weak, LS=large-strong).

	anger	fear	happiness	sadness
Mean Vp/d	25.2	17.3	22.9	5.8
Standard deviation	1.8	4.9	2.1	1.2
	SW	SS	LW	LS
Mean Vp/d	13.7	17.8	19.4	20.5
Standard deviation	2	1.8	2.6	1.5

Table 2.2 Results of the two-way repeated-measures ANOVA of emotion and fingerings on keystroke dynamics (as reflected by Vp/d).

	F	df	*p*	η_p^2
emotion	8.26	3,21	< 0.001	0.31
fingerings	4.05	3,21	< 0.001	0.16
emotion* fingerings	2.17	9,63	< 0.05	0.13

Table 2.3 Results of *post-hoc* Tukey tests on means of the four levels of emotion (A=anger, F=fear, H=happiness, S=sadness) and the four levels of fingerings (SW=small-weak, SS=small-strong, LW=large-weak, LS=large-strong).

	A vs. F	A vs. H	A vs. S	F vs. H	F vs. S	H vs. S
p	< 0.05	> 0.05	< 0.001	< 0.05	< 0.05	< 0.001
	SW vs. LS	SW vs. LW	SW vs. SS	LS vs. LW	LS vs. SS	LW vs. SS
p	< 0.01	< 0.01	< 0.05	> 0.05	> 0.05	> 0.05

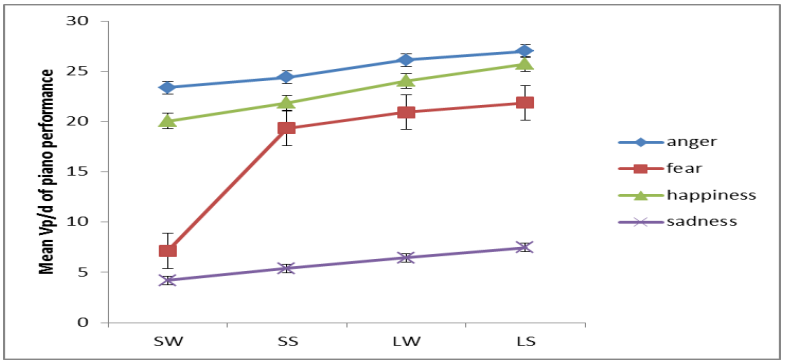


Figure 2.5 The interaction between emotions and fingerings (SW=small-weak, SS=small-strong, LW=large-weak, LS=large-strong) in terms of Vp/d in piano performance. Error bars represent the standard error of the mean.

2.3 Experiment 2: The speech experiment

2.3.1 Methods

Stimuli

The stimuli consist of six sentences divided into two sets (Tables 2.4, 2.5), with tones and vowels being the two variables. The purpose was to use the two variables to test the respective articulatory constraints on formant dynamics. According to the

reviews in section 2.1.2, tone 2 + tone 2 and tone 4 + tone 4 were used to create high articulatory pressure. Tone 3 + tone 1 was used to create low articulatory pressure. Meanwhile, a wide diphthong /au/ was used for long articulatory distance and a narrow diphthong /əu/ was used for short articulatory distance. *Cuilaoyao* and *Cuilouyou* are compound words denoting a person's name.

Table 2.4 The first set of the stimuli in which the numbers of the syllables represent the five lexical tones in Mandarin: 1 for H (High tone), 2 for R (Rising tone), 3 for L (Low tone), 4 for F (Falling tone), and 5 for N (Neutral tone).

	lao2[lau] work	*yao2[jau]* distant				
	lao3[lau] old	*yao1[jau]* waist	*nian4* read	*shu1* book	*qu4* aspect	*le5* particle
cui1 surname	*lao4[lau]* flood	*yao4[jau]* medicine				

IPA transcriptions for the key words *laoyao* are provided in brackets. Translation: *cuilaoyao* has gone to read a book.

Table 2.5 The second set of the stimuli in which the numbers of the syllables represent the five lexical tones in Mandarin: 1 for H (High tone), 2 for R (Rising tone), 3 for L (Low tone), 4 for F (Falling tone), and 5 for N (Neutral tone).

	lou2[ləu] building	*you2 [jəu]* oil				
cui1 surname	*lou3[ləu]* hug	*you1[jəu]* good	*nian4* read	*shu1* book	*qu4* aspect	*le5* particle
	lou4[ləu] drip	*you4[jəu]* right				

IPA transcriptions for the key words *louyou* are provided in brackets. Translation: *cuilouyou* has gone to read a book.

Measurement of formant dynamics

As reviewed in section 2.1.2, formant dynamics are an important factor reflecting the articulatory effort of speech production. Formant peak velocity, i.e., "the highest absolute value in the continuous velocity profile of the (formant) movement" (Cheng and Xu, 2013, p. 4488), and the displacement/amplitude of the formant movements are particularly related to articulatory effort [cf. Cheng and Xu (in press) for further discussion]. The peak velocity is measured in the following way (Xu and Wang, 2009, p. 506):

"Positive and negative extrema in the velocity curve correspond to the rising and falling ramps of each unidirectional pitch (formant) movement. A velocity curve was computed by taking the first derivative of an F0 (formant) curve after it has been smoothed by low-pass filtering it at 20 Hz with the Smooth command in Praat. Following Hertrich and Ackermann (1997), the velocity curve itself was not smoothed so as not to reduce the magnitude of peak velocity."

Figures 2.6 and 2.7 show the measurement points taken from F1 and F2 formant contours. This allows the calculation of the ratio of formant peak velocity (Vp) to maximum formant displacement (d), henceforth Vp/d. It reflects articulatory effort/vocal vigorousness (Cheng and Xu, 2013, in press). Similar to the piano experiment, $\text{Vp}=\frac{maximum\ formant\ displacement\ (d)}{time\ taken\ to\ reach\ the\ maximum\ displacement\ (t)}$, and so $\text{Vp/d} = \frac{d}{t} \times \frac{1}{d} = \frac{1}{t}$. The unit of displacement is Hz and the unit of time is sec.

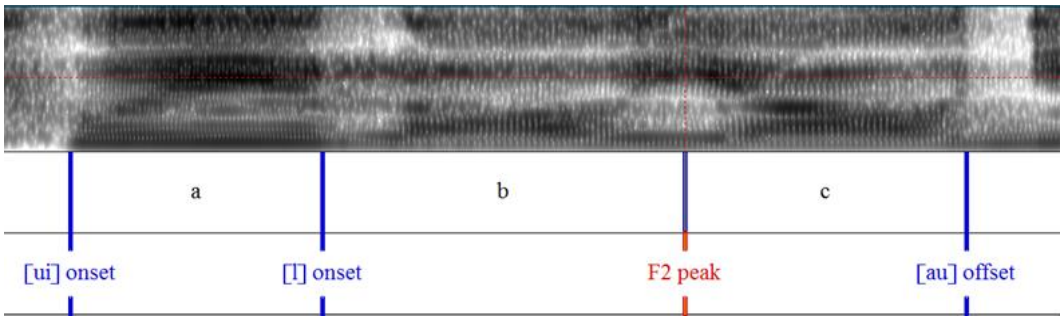


Figure 2.6 Syllable segmentation and labelling of the syllables /ui/ (a), /lau/ (b) and /jau/ (c) in the sentence "*Cui laoyao nian shu qu le*".

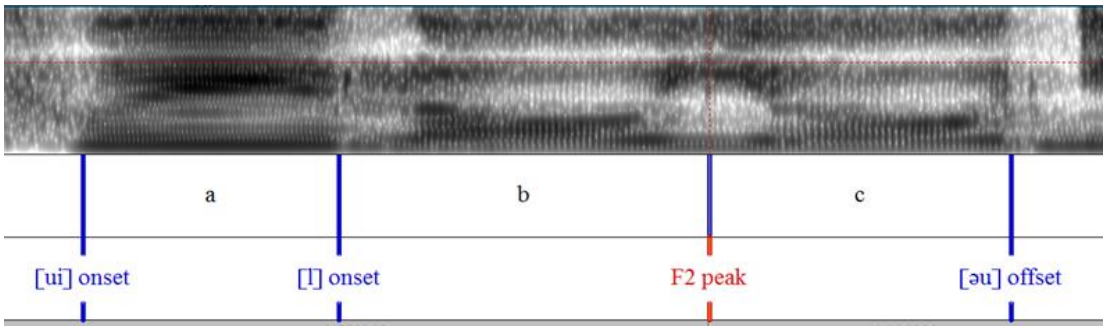


Figure 2.7 Syllable segmentation and labelling of the syllables /ui/ (a), /ləu/ (b) and /jəu/ (c) in the sentence "*Cui louyou nian shu qu le*".

Table 2.6 lists the values extracted from the measurement points for the calculation of Vp/d for F1 and F2.

Table 2.6 Values taken from the measurement points a, b, c for the calculation of Vp/d.

minF1a: F1 minimum in *a*	minF2a: F2 minimum in *b*
maxF1a: F1 maximum in *b*	maxF2a: F2 maximum in *a*
minF1b: F1 minimum in *b-c*	minF2b: F2 minimum in *c*
maxF1b: F1 maximum in *c*	maxF2b: F2 maximum in *b-c*
D1a: maxF1a – minF1a [F1 rising displacement]	D2a: maxF2a – minF2a [F2 falling displacement]
D1b: maxF1a – minF1b [F1 falling displacement]	D2b: maxF2b – minF2a [F2 rising displacement]
D1c: maxF1b – minF1b [F1 rising displacement]	D2c: maxF2b – minF2b [F2 falling displacement]
V1a: F1 peak rising velocity, in *a-b*	V2a: F2 peak falling velocity, in *a-b*
V1b: F1 peak falling velocity, in *a-b*	V2b: F2 peak rising velocity, in *a-b*
V1c: F1 peak rising velocity, in *b-c*	V2c: F2 peak falling velocity, in *b-c*

Subjects and procedure

Ten native Mandarin speakers without speech or hearing problems were recruited as subjects (5 females; Mean = 27 years, SD =2.5) via the University College London Psychology Pool. The recording session for each participant lasted for around half an hour. This experiment was approved by the Committee on Research Ethics at University College London. Voice portrayal/simulation method was used to induce

emotions, i.e., the participants were asked to imagine themselves in emotion-triggering scenarios when recording the sentences. This is because compared to other emotional speech induction methods (e.g., natural vocal expression), this method is more effective in obtaining relatively genuine emotional speech when experimental control is a key concern. Support for this method comes from the fact that natural emotional expression is often inherently involved with unintended portrayal and self-representation (Scherer, 2003). The recording was conducted in a sound-controlled booth. Participants were asked to record each sentence 3 times in four emotions: anger, fear, happiness and sadness, resulting in 10 (speakers) * 4 (emotions) * 3 (tones) * 2 (segments) * 3 (repetitions) = 720 tokens.

Similar to the first experiment, a follow-up perception validation test was conducted: twenty native speakers of Mandarin (11 females, Mean=23 years, SD=2.6) were asked to rate each emotion*tone*segment token on a scale of 1 to 5. 1 represents not at all angry/fearful/happy/sad while 5 represents very angry/fearful/happy/sad. The top eight ranked tokens (out of 30) for each emotion*tone*segment were selected. The mean score for each emotion*tone*segment was 4.16. ProsodyPro and FormantPro scripts (Xu, 2014) running under Praat (Boersma and Weenink, 2013) were used for data analyses.

2.3.2 Results

The means of Vp/d of all measurement points are represented in Table 2.7. A three-way repeated measures ANOVA showed that among the three factors (emotion, tone and segments), emotion was the only factor exerting a significant impact on the value of Vp/d. The interaction between emotion, tone and segments was non-

significant. However, the interactions between emotion and tone and that between emotion and segments were significant (Table 2.8).

Table 2.7 Mean Vp/d of the four levels of emotion (A=anger, F=fear, H=happiness, S=sadness) and the four levels of articulatory constraints (SS=small segmental distance /əu/, ST=small tonal pressure T3+T1, LS=large segmental distance /au/, LT=large tonal pressure T2+T2/T4+T4).

	anger	fear	happiness	sadness
mean Vp/d	37.3	41.5	43.6	24
standard deviation	6.2	5.5	5.6	4.9
	SS	ST	LS	LT
mean Vp/d	34	33.6	38.5	39.8
standard deviation	5.9	4.8	5.1	5

Table 2.8 Results of the three-way repeated-measures ANOVA on articulation dynamics (as reflected by Vp/d).

significant effects	F	df	p	η_p^2
emotion	5.22	3,21	< 0.01	0.25
emotion*tone	2.39	6,42	< 0.05	0.11
emotion*segments	3.08	3,21	< 0.05	0.12

Post-hoc Tukey tests showed more details: sadness had significantly ($p < 0.05$) the lowest Vp/d value compared with the other three emotions. Happiness had the highest dynamics followed by fear and anger, but the differences between the three were non-significant.

The interaction between emotion and tonal pressure was significant. As shown in Figure 2.8, the Vp/d of all emotions was higher in tonal combinations of large articulatory constraints (i .e., T2 + T2 and T4 + T4) than the Vp/d in those of small articulatory constraints (T3 + T1). This is the most obvious in the case of anger where T2 + T2 and T4 + T4 made the Vp/d of anger become closer to that of fear and happiness. *Post-hoc* Tukey tests showed that the differences between anger and fear plus the differences between anger and happiness were non-significant under the T2 + T2 and T4 + T4 conditions. In contrast, under the T3 + T1 condition, the differences between anger and fear plus the differences between anger and happiness were significant (both *p*s < 0.05). In addition, for fear, happiness and sadness, the Vp/d did not differ significantly between the two tonal conditions. Therefore, anger was more affected by tonal variation than the other three emotions.

The interaction between emotion and segments was also significant. Figure 2.8 shows Vp/d was overall higher in the wide diphthong condition than in the narrow condition. The interaction was the most obvious in the case of anger because it was almost as high as fear and happiness with regard to Vp/d in the wide diphthong condition. *Post-hoc* Tukey tests showed the differences between anger and fear plus the differences between anger and happiness were non-significant in the wide diphthong condition. The differences were significant (both *p*s< 0.05), however, under the narrow diphthong condition. Moreover, for fear, happiness and sadness, the Vp/d did not differ significantly between the two segmental conditions. Therefore, similar to above, anger was more influenced by segmental distance variation than the other three emotions.

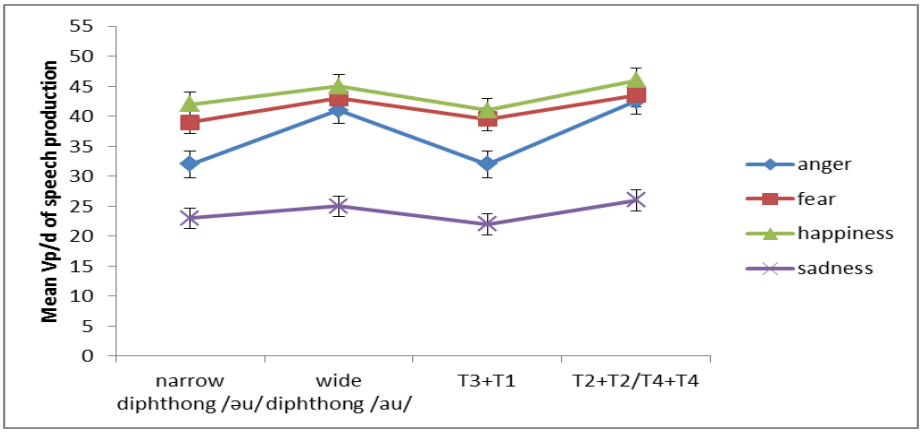


Figure 2.8 The mean Vp/d of narrow diphthong, wide diphthong, T3 + T1 and T2 + T2/T4 + T4 in the four types of emotional speech (anger, fear, happiness, and sadness). Error bars represent the standard error of the mean.

2.4 Comparisons between the results of the piano and speech experiment

To directly compare the results of the piano and speech experiments, a MANOVA test was conducted: the within-subjects independent variables were emotion (four levels: anger, fear, happiness and sadness) and physical constraint (two levels: large hand span/articulatory distance; small hand span/articulatory distance) while the between-subjects independent variable was group (two levels: pianists and speakers). The dependent variables were Vp/d (speakers) and Vp/d (pianists). Using Pillai's trace, there was a significant difference between pianists and speakers ($F_{(8, 7)} = 13.78$, $p < 0.01$). The following univariate ANOVAs showed that the group differences between pianists and speakers were significant across most conditions: anger-large ($F_{(1, 14)} = 14.92$, $p < 0.01$), anger-small ($F_{(1, 14)} = 16.23$, $p < 0.01$), happiness-small ($F_{(1, 14)} = 15.61$, $p < 0.01$), fear-large ($F_{(1, 14)} = 14.95$, $p < 0.01$), fear-small ($F_{(1, 14)} = 18.09$, $p < 0.01$), sadness-large ($F_{(1, 14)} = 15.93$, $p < 0.01$). In the happiness-large and sadness-small conditions, the group difference was non-significant (although speech production still had higher Vp/d than that of piano performance). The results suggest

on the whole, piano performance has significantly different (i.e., lower) dynamics than speech production.

2.5 Discussion and conclusion

2.5.1 Similarities between affective piano performance and speech production

The results showed that firstly, anger in piano performance generated the highest dynamics irrespective of fingerings; in speech production, it was also relatively high in dynamics although it interacted with articulatory constraints (more discussions on the interaction are offered in the following section). This is in line with previous reports that anger in music performance and speech production is generally linked to high intensity and great energy (cf. Juslin and Sloboda, 2013). Physiologically, the high dynamics of anger can be associated with high levels of cardiovascular activities such as high heart rate (Rainville et al., 2006), fast/deep breathing (Boiten et al., 1994), increases in diastolic pressure and activated baroreceptor mechanisms (Schwartz et al., 1981). Evolutionarily, anger originates from natural selection pressure on animals (Darwin, 1872): anger induces in humans the inclination to fight or attack whatever that threatens survival and well-being. As a result, anger is proposed to be associated with large body size projection (Morton, 1977; Xu et al., 2013a, 2013b) to scare off enemies. Hence anger should be linked to high dynamics which can be reflected by high physical or vocal effort to show great strength and energy (Xu et al., 2013a). The results of this study support this prediction by demonstrating that greater finger force and articulatory effort were generated respectively in piano performance and speech production in the context of anger.

Secondly, happiness triggered the highest dynamics for speech production and second highest dynamics for piano performance, irrespective of fingerings or

articulatory constraints. The results are in line with previous reports that in music performance, happiness is always associated with faster tempo and higher intensity (Gabrielsson, 1995; Widmer and Goebl, 2004; Zanon and De Poli, 2003a, 2003b); happy speech is reported to have high values in many acoustic dimensions such as pitch, pitch range, intensity (Scherer, 2003; Ververidis and Kotropoulos, 2006), speech rate and formant shift (Xu et al., 2013a). Similar to anger, the physiological reason for high dynamics of happiness is often linked to increases in heart rate, blood pressure, breathing pattern (Boiten et al., 1994; Rainville et al., 2006), all of which can contribute to greater physical or vocal force in music performance or speech production. From an evolutionary perspective, happiness can be a useful strategy for attracting mates (Darwin, 1872). Therefore, it is beneficial for sound signalers to produce highly vigorous (i.e., dynamic) sounds so as to be audible to potential mates (Xu et al., 2013a). Hence, the results are also consistent with the evolutionary account.

Thirdly, fear in both piano performance and speech production was linked to significantly higher dynamics than sadness; particularly in speech production fear did not differ significantly from anger/happiness. This might seem somewhat unexpected particularly in terms of music, given that fear in music performance is generally associated with soft playing similar to sadness (cf. Juslin and Sloboda, 2013). In terms of speech production, however, fear has already been found to show high dynamics (Xu et al., 2013a), which is consistent with the view that evolutionarily, fear can be a defensive emotion (LeDoux, 1996), evidenced from animal alarm calls as a useful antipredator defensive strategy across many species for the sake of group survival (Caro, 2005). To serve this purpose, alarm calls should be reasonably high in dynamics (i.e., vigorousness). Similarly, production of musical

excerpts of fear could also be highly dynamic, analogous to human fearful speech or animal alarm calls.

Fourthly, sadness always generated the lowest dynamics for both piano and speech performance regardless of fingerings or articulatory constraints. This finding is in line with previous research: sad music and speech are generally low in acoustic cues such as intensity, F_0, F_0 range and duration (Juslin and Laukka, 2003; Laukka et al., 2005; Patel, 2008). This is mainly because sadness is located at the opposite end of happiness in terms of valence and arousal: it is a lowly aroused negative emotion because of its association with reduced physiological energy and arousal level, sometimes leading to affective pathology such as depression or anhedonia (cf. Huron, 2011). Evolutionarily, such low dynamics of sadness indicate physical and social withdrawal and a tendency for the sound signaller to beg for sympathy (Shaver et al., 1987). Hence usually low motor effort is involved in expression of sadness either through music or speech. It is worth mentioning sad speech can be split into two categories: depressed sadness and mourning sadness (Scherer, 1979), the former being characterized by low vocal energy while the latter by high vocal energy. In this study, it was the depressed sadness that was used and hence the resulting formant dynamics were low, reflecting decreased articulatory effort due to the sluggishness of articulatory muscles in sad speech (Kienast and Sendlmeier, 2000).

2.5.2 Differences between affective piano performance and speech production

The results also showed significant differences between the two domains. The most notable difference is that speech production on the whole had higher dynamics than piano performance across almost all conditions. This is consistent with previous

studies on comparisons between speech articulatory movements and limb movements (Gentil and Tournier, 1998; Ito et al., 2004; Loucks et al., 2010). Although those studies did not investigate movements in the context of affective piano performance or speech production, the general biophysical mechanisms of fingers and speech articulators apply to this study. More specifically, it was found that compared with fingers or arms, speech articulators in general produce faster velocity (Loucks et al., 2010; Ito et al., 2004) and greater force (Gentil and Tournier, 1998; Loucks et al., 2010). The reasons probably lie in the biomechanical differences between speech articulators and fingers: compared with speech articulators, fingers are associated with more intervening factors (e.g., long tendons, joints and muscle mass between muscle fibers and skeletal joints) that prevent finger muscles from contracting as fast as speech articulatory muscles (Gentil and Tournier, 1998). It has also been reported that oral-facial muscles are associated with fast-twitch fibers and motor protein such as myosin which enable fast acceleration and rapid speech in order to meet different levels of speech demand (Burke, 1981; Williams and Warwick, 1980). Therefore, in this study the dynamics of affective speech production (as reflected by articulatory effort) and piano performance (as reflected by finger force) were different from each other due to the biomechanical distinctions.

In addition, the results also demonstrated that the interaction between emotion and physical constraints in piano performance was different from that in speech production. In piano performance (Figure 2.5), only fear interacted with physical constraints (i.e., fingerings); in speech production (Figure 2.8), only anger interacted with physical constraints (i.e., articulatory constraints). The reasons could be attributed to the differences in the extent of acoustic stability of music performance and speech production in different emotions.

Firstly, in music performance, anger, happiness and sadness are associated with relatively consistent acoustic patterns (Juslin and Sloboda, 2013), i.e., anger and happiness are always fast and loud to convey high energy and arousal while sadness is always slow and quiet to convey low energy and arousal. Fear, in contrast, is linked to highly variable acoustic patterns especially in terms of tempo and intensity (Madison, 2000; Juslin and Madison, 1999; Bernays and Traube, 2014; Juslin and Sloboda, 2013) so as to convey the unstable psychological state under the influence of fear, e.g., scattered notes with pauses between musical phrases and sharp contrasts between intensity are often used to express fear (Madison, 2000). This could further imply there may not be a consistent pattern of finger force under the condition of fear. Hence, other factors such as fingerings are highly likely to interact with fear to generate different kinematic dynamics in piano performance.

On the other hand, fearful speech shown in this study always had high formant dynamics regardless of articulatory constraints. This is likely to be associated with duration: under the condition of fear, the mean duration of the segmented syllables as shown in Figures 2.6 and 2.7 was 555.6ms. This was not significantly different from the mean duration of the segmented syllables under the condition of happiness (mean=546.6ms) which had the highest dynamics. Moreover, the duration of the segmented syllables in fear was significantly ($p < 0.05$) shorter than that in anger (mean = 601.1 ms) and sadness (mean = 638.2 ms). In addition, the difference in duration between large and small articulatory constraints was non-significant under the condition of fear. Similar findings have been reported that fear is often produced with fast speech rate that is likely to trigger vowel undershoot [i.e., an articulatory phenomenon where the canonical phonetic forms of speech sounds fail to be reached because of the articulatory impact of surrounding segments (Lindblom, 1963)] and

segmental reduction (Kienast and Sendlmeier, 2000; Paeschke et al., 1999). Shorter duration is highly likely to trigger great articulatory effort according to the report of studies on articulatory movement (Adams et al., 1993; Edwards et al., 1991; Munhall et al., 1985; Ostry and Munhall, 1985; Perkell et al., 2002). Therefore, the relatively stable acoustic pattern (i.e., duration) of fearful speech could make it less likely to interact with other factors such as articulatory constraints.

Secondly, this study showed that only angry speech significantly interacted with articulatory constraints: the formant dynamics were significantly higher in large articulatory constraints than those in small articulatory constraints. Again this can be linked to duration. A closer look at the data reveals that the duration of angry speech was significantly ($p < 0.05$) shorter under the condition of large articulatory constraints than the condition of small articulatory constraints. It has been reported (Cheng and Xu, 2013) that when time is short for the articulatory execution of segments with large articulatory constraints, muscles have to contract faster (i.e., with stronger articulatory effort) than when small articulatory constraints are involved in order to reach the tonal and segmental targets. This was reflected in the high formant dynamics under the condition of large articulatory constraints in this study. In addition, the result is also consistent with the finding that anger is often more variable in duration compared with the other three emotions (happiness, fear and sadness): it can be slow because of the need to be precise and clear in articulation (Kienast and Sendlmeier, 2000; Paeschke et al., 1999) so as to project big body size to threaten away enemies (Xu et al., 2013a, 2013b); it can also be fast in speech rate (Scherer, 2003) especially in female speakers to reflect the highly aroused and variable psychological state under the influence of anger. Hence, it is

the relatively high variability in duration that makes angry speech more prone to interact with external factors such as articulatory constraints.

All in all, the results on the one hand showed similar tendencies in dynamics of affective speech production and piano performance, which can be explained from a bio-evolutionary perspective. On the other hand, different interaction patterns were found: physical constraints interacted only with fear in piano performance while in speech production, only with anger. This suggests that the more variable an emotion is in acoustic features, the more likely it is to interact in production with external factors such as fingerings or articulatory constraints in terms of dynamics. In addition, speech production on the whole had higher dynamics than piano performance, which could be due to the bio-mechanical differences between speech articulators and fingers. Therefore, this is the first study to quantitatively demonstrate the importance of considering motor mechanisms such as dynamics (i.e., finger force and articulatory effort) together with physical constraints (i.e., fingerings and articulatory constraints) in examining the similarities and differences between affective music performance and speech production. In a nutshell, focusing on the motor mechanisms of affective music performance and speech production could further enhance our understanding of the relations between music and speech.

Chapter 3

Emotional connotations of musical timbre of isolated instrument sounds through the perspective of affective speech: Behavioural and ERP Evidence

3.1 Introduction

Music and speech are primary platforms for humans to communicate emotions (Buck, 1984). A considerable amount of studies have shown that affective music and speech are similar in many psychoacoustic dimensions: pitch, intensity and duration have been given extensive attention over a long period in terms of their cross-domain similarities (cf. Juslin and Laukka, 2003). Timbre is another important acoustic dimension but is not as well researched as pitch, intensity and duration (Eerola et al., 2012; Holmes, 2011). Only recently has timbre, especially musical timbre, attracted a reasonable amount of scholarly interest. This chapter will further explore musical timbre, particularly in terms of the emotional connotations of musical timbre of isolated instrument sounds, with affective speech as a reference. The reasons will be provided below.

3.1.1 Background on musical timbre

Timbre is a multidimensional auditory event enabling listeners to distinguish between sounds that have equal pitch, loudness and duration (Giordano and McAdams, 2010). Auditory processing of timbre requires perceptual integration of spectral and temporal dimensions (Griffiths and Warren, 2004). The significance of timbre in auditory processing is evidenced from the fact that even in infancy, humans can differentiate and memorize different types of timbre (Trehub et al., 1990). In

music theory, timbre is an effective platform for conveying composers' underlying intentions and inducing emotions from listeners (Boulez, 1987; Gabrielsson, 2001). Early empirical evidence for the association of timbre with emotion can be found in Scherer and Oshinsky (1977) where tone sequences were manipulated in tempo, pitch, intensity as well as timbre (i.e., spectral filtering and envelope manipulation). The systematic change in those acoustic dimensions led to listeners' ratings of different emotions. Studies have also shown that listeners could distinguish emotion categories from very short musical excerpts, e.g., 250ms long (Peretz et al., 1998), 400ms long (Krumhansl, 2010) or 1s long (Bigand et al., 2005). This suggests timbre could be more immediate to the recognition of emotion than other music cues which usually take longer to process (Eerola et al., 2012).

Studies on polyphonic musical timbre (i.e., timbre of more than one instrument) (e.g., Alluri and Toivianen, 2010) show that the arousal dimension of emotion is strongly correlated with the high-low frequency energy ratio of spectrums. Cross-cultural studies on music perception have shown that western listeners tend to perceive flute in Hindustani music as conveying peacefulness while strings as implying anger (Balkwill and Thompson, 1999). In western music, on the other hand, flute is often associated with sadness (Balkwill et al., 2004). Recently, an in-depth study of monophonic timbre (i.e., timbre of one instrument) (Eerola et al., 2012) investigated emotional connotations of isolated instrument sounds through a series of perception experiments. It was found that affective dimensions (i.e., arousal and valence) of the instrument sounds were mainly determined by spectral (high-low frequency ratio), temporal (attack slope) and spectro-temporal (spectral flux) parameters. Listeners' consistent ratings of valence and energy arousal of the instrument sounds across experiments further indicate that timbre is a primary cue of conveying musical

emotions. ERP experiments on musical timbre (Goydke et al., 2004) show that the MMN (mismatch negativity) could be triggered by the violin playing the same melody with different emotions (happiness and sadness) as standards or deviants, hence leading to the conclusion that variation in musical timbre can communicate emotion. A follow-up ERP study (Spreckelmeyer et al., 2013) further extended this experiment by incorporating more timbral variations of each emotion (happiness and sadness) as standards. The results showed MMN can still be elicited under pre-attentive condition, suggesting that in spite of the variance, the standards were still grouped together as a unified emotional entity (Spreckelmeyer et al., 2013).

3.1.2 Problems with previous studies on musical timbre

The studies reviewed above demonstrate a strong connection between musical timbre and emotion. Nevertheless, a problem common to almost all of the above studies is that timbre was not tested as an independent acoustic cue free from variations in other acoustic cues such as pitch, duration and intensity, i.e., acoustic features other than timbre were not strictly controlled in those studies. The study by Eerola et al. (2012) (reviewed above) has a relatively stricter control, but special musical effects (e.g., vibrato, flutter) were not filtered out. It is known that effects like vibrato and flutter involve modulations in pitch and intensity (Olson, 2003). Therefore, it is not clear whether it was timbre alone or the combination of many acoustic features that contributed to the perceptual judgement of emotion reported in those studies. Hence, greater effort with a much more focused attention on timbre alone would be necessary to advance our understanding of the emotional connotations of musical timbre.

In addition, to our knowledge there is not enough research that directly compares emotional connotations of the timbre of isolated instrument sounds with human affective speech. We believe it is worth further exploring the relations between the two domains. This is because firstly, human voice, a crucial platform for conveying speaker's emotion and attitude (Banse and Scherer, 1996; Gobl and Ní Chasaide, 2003), has long been compared to musical instruments. Richard Wagner, the famous composer, believed that the oldest and most natural embodiment of musical instruments is the human voice (cf. Watson, 1991). Similarly, Stendhal, the famous novelist, once commented that only the musical instruments that approximate the human voice can be truly pleasing to the human ear (cf. Watson, 1991). String instruments such as the violin and the guitar are classic examples of the approximation of musical instruments to human vocal expressions (Askenfelt, 1991).

Secondly, there is neuropsychological evidence showing that string instrument timbre and the human voice elicit similar ERP responses (Levy et al., 2003). Nevertheless, in that study the voice stimuli were sung vowels devoid of linguistic meaning and also emotion was not included as a factor. Hence, it is still unknown whether human affective speech (with meaningful linguistic information) could trigger similar or different ERP responses compared with those of musical instrument timbre. Brain imaging reports also show that during the perception of musical timbre, evidence for cognitive processing of emotion was found in the P200 time window, as suggested by the additional anterior cingulate cortex (ACC) activities (Meyer et al., 2006). This evidence would be much stronger if human affective speech was included as a factor for comparison with musical instrument timbre.

Relatively more direct comparisons between the two domains have used affective priming paradigm with visually presented words as primes or targets (Goerlich et al., 2012; Painter and Koelsch, 2011; Steinbeis and Koelsch, 2011). Affective priming refers to the phenomenon where the processing speed of an affective stimulus (e.g., the word "happy") becomes faster when preceded by stimulus of the same affective category (e.g., the word "sunny") than that of a different category (e.g., the word "boring") (Klauer and Musch, 2003). The N400 response has been found a primary reflector of the affective priming effect. Originally found in studies on semantic incongruity (Kutas and Hillyard, 1980; for a recent review, see Kutas and Federmeier, 2011), the N400 effect was later found to be elicited in a variety of domains such as environmental sounds (Frey et al., 2014; Orgs et al., 2006, 2007, 2008; van Petten and Rheinfelder, 1995), odours (Grigor et al., 1999), pictures (Hamm et al., 2002), affective speech (Paulmann and Pell, 2010; Schirmer et al., 2002) and music (Steinbeis and Koelsch, 2011; Painter and Koelsch, 2011).

In particular, the N400 discovered in music studies suggests that music can trigger meaning, emotional meaning in particular. Steinbeis and Koelsch (2011) showed that musical instrument timbre could communicate emotion to musically trained and untrained listeners. Short musical chords (800 ms) subjectively rated as pleasant or unpleasant were used as primes followed by visually presented words congruent/incongruent with the emotional valence of the chords. Words emotionally congruent with the chords (e.g., pleasant sounding chords followed by the word "beauty") triggered smaller N400 amplitude than words emotionally incongruent with the chords (e.g., pleasant sounding chords followed by the word "anger"). In another study, Painter and Koelsch (2011) focused on the semantic meaning of out-of-context music sounds while controlling for emotion. A larger N400 amplitude

was triggered by semantically incongruent sound-word pairs than that of semantically congruent sound-word pairs, under the condition of participants' active evaluation. Nevertheless, in the aforementioned studies, timbre was not independent of the variation in other acoustic features such as pitch or duration, and so it is not clear if it was timbre or the combination of acoustic features that contributed to the N400 effect. In addition, the focus of the aforementioned studies was on the semantic level of words; larger linguistic unit such as speech was not investigated. Moreover, the linguistic stimuli were visually presented, rather than auditorily presented. Given the close acoustic connections between affective music and speech (Juslin and Luakka, 2003), it is worth exploring the relations between affective speech and musical timbre from an auditory perspective.

3.1.3 The present study

In summary, the above reviews suggest that there is a strong link between musical timbre and emotion. However, previous behavioural and neuropsychological studies have ignored the following issues: 1) proper control of other musical features (e.g., pitch, duration, intensity); 2) musical timbre of isolated instrument sounds in relation to affective human speech from an auditory perspective. We believe these issues are vital for a proper and more enhanced understanding of the emotional meaning of musical timbre, particularly in terms of its connection with affective speech. Hence, this study aimed to explore emotional connotations of musical timbre of isolated instrument sounds through the perspective of affective speech, with a strict control of other musical features. More specifically, we tested whether musical timbre of isolated instrument sounds *alone* can have emotional connotations (i.e., whether timbre can communicate emotion). Emotion, in this study, was represented by affective speech. This means affective speech was used as a reference to which

musical instrument timbre was compared (the reason has been provided in previous paragraphs), with the goal of testing whether a single acoustic dimension of isolated musical instrument sounds (i.e., timbre) is capable of communicating emotion in a way similar to affective speech. In addition, as introduced in Chapter 1, another aim of this study is to test whether musical instrument timbre has acoustic characteristics that convey emotion in the same direction as predicted by the body-size projection theory on affective speech.

A behavioural and two ERP experiments were carried out to achieve the aforementioned aims. The behavioural experiment tested whether emotions conveyed by musical timbre of isolated instrument sounds would have similar timbral patterns as those of affective speech. The first ERP experiment focused on the P200 and LPC (late positive complex) components which are primary indicators of cognitive processing of emotional stimuli (cf. Paulmann et al., 2013). We tested whether the ERP patterns elicited by instrument sounds would be similar to those of affective speech. The second ERP experiment focused on the N400 using the affective priming paradigm, with the aim to test whether emotionally incongruent instrument-speech pairs would elicit larger N400 than emotionally congruent instrument-speech pairs. The second ERP experiment is a logical extension of the first ERP experiment: The first one tested the brain's processing of musical instrument sounds and affective speech separately, while the second one tested the brain's response when the two categories of stimuli were presented via priming (i.e. more closely together) through the lens of N400. As reviewed above, N400 has been found to indicate emotional meaning in music through cross-modal (e.g. music and words) priming paradigms (Steinbeis and Koelsch, 2011). Hence, the second ERP experiment more directly addressed the research aim of this study: the emotional

connotations/meaning of musical instrument timbre through the perspective of affective speech. In addition, it is not unusual to examine different ERP components such as P200, LPC and N400 in one study (e.g., Dunn et al., 1998; Iakimova et al., 2009), particularly in the field of emotion and meaning (e.g., Kanske and Kotz, 2007; Kanske et al., 2011).

3.2 The behavioural experiment

The aim of the behavioural experiment was twofold: first, to compare acoustic characteristics of affective speech timbre (voice quality) and musical timbre; second, to select affective speech and musical instrumental stimuli for the follow-up ERP experiments.

3.2.1 Methods

Participants

Twenty native speakers of Mandarin (10 females, age *M* =28.3, *SD* = 5.2) without music training background were recruited as participants.

Stimuli

The speech database includes a pre-recorded Mandarin sentence (*Cui luya nian shu qu le*, meaning *Cui luya* has gone to read a book) produced in three emotions (anger, happiness, and sadness) by 8 native Mandarin speakers (4 females, age *M* = 25.3, *SD* = 2.1, different from the 20 participants for this behavioural experiment), with each rendition of the sentence per emotion repeated three times. Therefore there were 8 (speakers) * 3 (emotions) * 1 (sentence) * 3 (repetitions) = 72 trials for the speech experiment. The music database was from McGill University Master Samples (MUMS) (Opolko and Wapnick, 2006). This database includes sounds of almost all

instruments (110 sounds altogether). Following Eerola et al. (2012), all the instrumental sounds were equal in pitch (D ♯ 4). For the purpose of this study, the sounds were also made equal in duration (1s). Moreover, the loudness of all the sound samples was adjusted in a way to ensure a perception of equal-loudness. In addition, sounds that have special effect (e.g., vibrato) were removed from the dataset. The purpose of this was to guarantee that other than timbre, all the rest of the acoustic features of the music stimuli remained perceptually the same.

Procedure

The participants were instructed to complete a speech and a music task. For the speech task, they were asked to listen to the 72 sentences in the speech database and rate each sentence per emotion on a 1-5 scale which indicated the intensity of the emotion (1 meant very weak; 5 meant very strong). The top 4 rated sentences of each emotion category were selected for the following behavioural analyses and ERP experiments. The mean score for all selected sentences in each emotion category was above 4.5. The acoustic features of the selected speech items were: anger (pitch: 297.74Hz; duration: 1359ms; intensity: 77dB); happiness (pitch: 260.48Hz; duration: 1338ms; intensity: 68.18dB); sadness (pitch: 199.36Hz; duration: 2091ms; intensity: 54.85dB). For each emotion, the speech items were from 2 females and 2 males (all different speakers). For the music task, the participants were asked to listen to the 110 musical instrument sounds and categorize each sound into one of the following categories: anger, happiness, sadness and neutral (no obvious emotion). Then each sound (except the sounds categorized as neutral) was rated on a 1-5 scale which indicated the intensity of the emotion (1 meant very weak; 5 meant very strong). The top 4 rated sounds of each emotion category (anger, happiness and sadness) were selected for the following behavioural analyses and ERP experiments. The mean

score for all selected sounds in each emotion category was above 4.3. Angry instruments selected were: cornet, alto shawm, crumhorn and saxophone. Happy instruments selected were: harpsichord, marimba, vibraphone and piano. Sad instruments selected were: violin, bassoon, flute and oboe.

The reason why only 4 tokens from each emotion category were selected is that it is necessary to ensure the stimuli were highly representative of each emotion. This is because acted affective speech (as in this study) sometimes cannot convey the targeted emotion satisfactorily (cf. Scherer, 2003). A stimulus size larger than 4 in this study would involve sound tokens not well produced for each emotion. With regard to music stimuli, listeners' judgment varied considerably as to the musical instruments for each emotion, and the top 4 instruments for each emotion were those that achieved high emotion validity scores (above 4). All the remaining instrument sounds for each emotion did not achieve an average score of more than 3, which consequently could not qualify as representatives of the targeted emotions. As a result, this study is limited in stimulus size and future research could include more sound tokens per emotion for wider generalizability.

Following Eerola et al. (2012), 6 timbral features (Table 3.1) were selected: attack slope and spectral centroid corresponding to temporal features; ratio of high-low frequency energy, spectral skewness and spectral regularity corresponding to spectral features; and spectral flux corresponding to spectro-temporal features. They are all important timbral features which could contribute to the emotional connotations of timbre (Eerola, et al., 2012). The features were extracted from the speech and instrument stimuli using the MIR toolbox (cf. MIRtoolbox User's Guide 1.6.1 by Lartillot (2014) for further technical details in computing each timbral feature). It is worth pointing out that timbre in music may correspond only partly with timbre (i.e.,

voice quality) in speech because in speech literature, voice quality is still a relatively vague term that lacks a precise definition. Nevertheless, the acoustic features (e.g., attack slope, spectral centroid and spectral skewness, etc.) selected in this study have been shown to be important parameters for examining affective voice quality by several studies (e.g., Banse and Scherer, 1996; Goudbeek and Scherer, 2010; Laukka et al., 2005; Xu et al., 2013a, 2013b).

Table 3.1 Definitions of the 6 timbral features selected for this study [cf. Eerola et al. (2012) and Lartillot (2014)].

Attack Slope	Slope of the attack portion of the sound
Spectral Centroid	Geometric center of the spectrum (McAdams et al., 1995)
Spectral Skewness	Symmetry of the spectral distribution
Spectral Regularity	Degree of uniformity of the successive peaks of the spectrum, also called Spectral Smoothness (McAdams et al., 1999)
Ratio of high-low frequency energy	High-low spectral energy ratio (Juslin, 2000)
Spectral Flux	Change between the consecutive spectral frames (McAdams et al., 1995)

3.2.2 Results

Figure 3.1 displays the means and standard deviations of the six timbral features of speech and instruments in the three emotions. Table 2 further summarizes the patterns of the three emotions (A=anger, H=happiness, S=sadness) with regard to the six timbral features of speech and musical instruments respectively. It can be observed that the patterns of emotions were similar for speech and instruments across all of the six timbral features, especially in terms of anger and happiness: happy speech and instruments had higher values than angry speech and instruments (H > A) in terms of spectral skewness and spectral flux; with regard to attack slope, spectral centroid, regularity and high-low frequency energy ratio, angry speech and instruments had higher value than happy speech and instruments (A > H). Sad speech and instruments had the lowest spectral centroid, highest spectral skewness and regularity among the three emotions. Nevertheless, the patterns of sadness were not consistent between speech and instruments with regard to attack slope, high-low frequency energy ratio and spectral flux.

To test the group differences between speech and musical instruments, three MANOVAs were carried out on the top 4 rated speech and instruments in the conditions of anger, happiness and sadness respectively. The dependent variables were the six timbral features: attack slope, spectral centroid, spectral skewness, spectral regularity, ratio of high-low frequency energy and spectral flux. The independent variable was group (two levels: affective speech and musical instruments). The results show that in anger, happiness and sadness, the differences between speech and musical instruments were non-significant (anger: $F_{(6,1)} = 31.03$, $p = 0.137$; happiness: $F_{(6,1)} = 19.13$, $p = 0.173$; sadness: $F_{(6,1)} = 5.485$, $p = 0.316$).

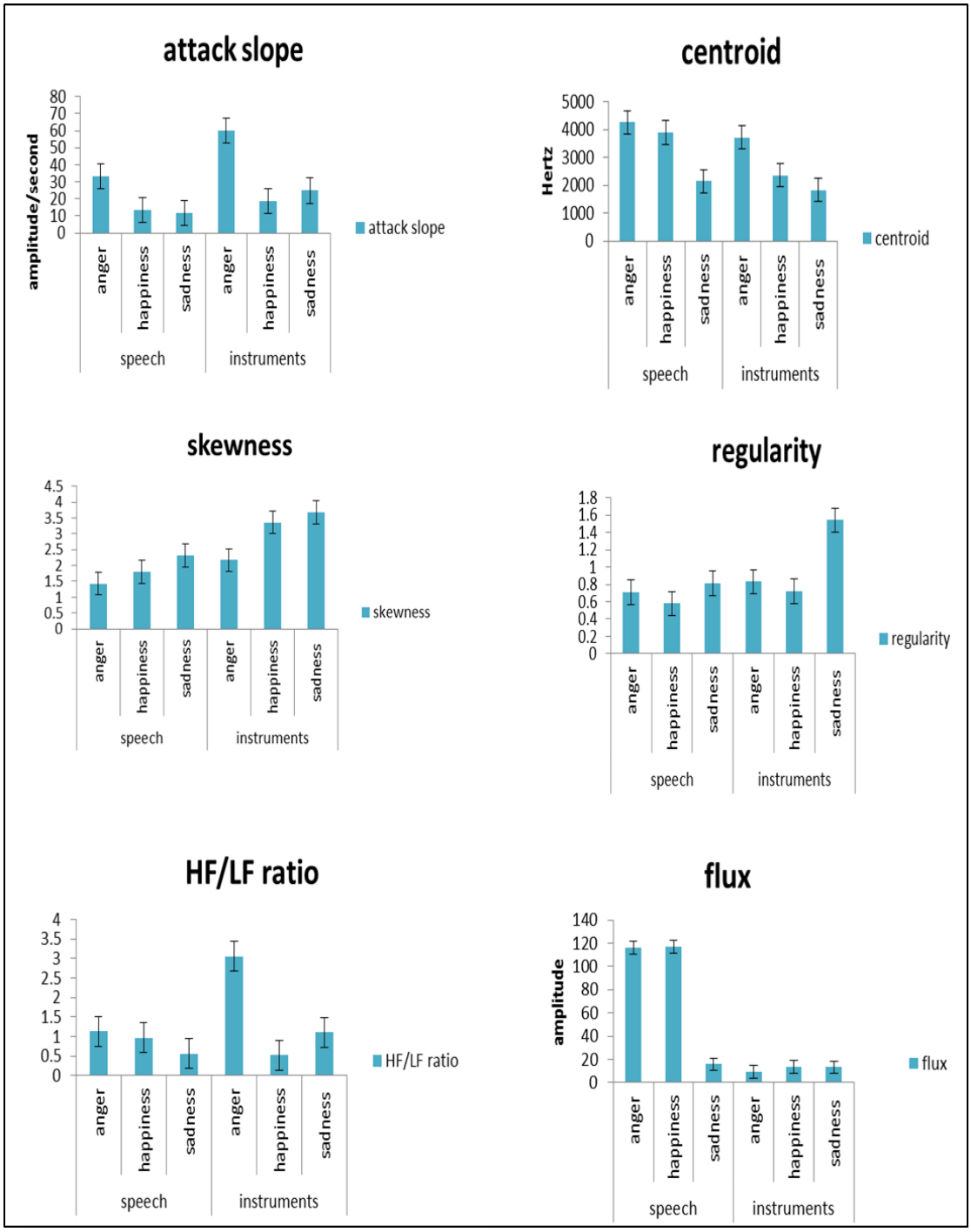


Figure 3.1 Means of the six timbral features of speech and instruments in the three conditions of emotion. Error bars represent the standard error of the mean.

Table 3.2 Patterns of the three emotions (A=anger, H=happiness, S=sadness) with regard to the six timbral features of speech and musical instruments respectively (significant comparisons are indicated in bold in the second line of each stimulus type, $p<0.017$).

	attack slope	centroid	skewness	regularity	HF/LF ratio	spectral flux
Speech	A>H>S (**A>H**, **A>S**, H>S)	A>H>S (A>H, **A>S**, **H>S**)	S>H>A (S>H, **S>A**, H>A)	S>A>H (S>A, **S>H**, A>H)	A>H>S (A>H, **A>S**, H>S)	H>A>S (H>A, **H>S**, **A>S**)
Instruments	A>S>H (**A>S**, **A>H**, S>H)	A>H>S (**A>H**, **A>S**, H>S)	S>H>A (S>H, **S>A**, **H>A**)	S>A>H (**S>A**, **S>H**, A>H)	A>S>H (**A>S**, **A>H**, S>H)	H>S>A (H>S, H>A, S>A)

3.3 ERP experiments

3.3.1 Methods

Participants and stimuli

Sixteen native speakers of Mandarin Chinese were recruited to take part in the ERP experiment 1 and another sixteen speakers of Mandarin Chinese for the ERP experiment 2 (8 females for each experiment, age M=23, SD=1.8 of the speakers for the first experiment; age M=26, SD=2.1 of the speakers for the second experiment). For each experiment, data from one participant was discarded due to excessive muscle artefacts. The participants reported no hearing or speech impairments. The experiments were approved by UCL research ethical committee. The stimuli for the ERP experiments were the top four rated speech and musical instruments from the behavioural experiment (cf. section 3.2 for details).

Procedure

Two ERP experiments were carried out on two separate days.

Experiment 1

The first experiment aimed to separately test the effect of emotional speech and musical instruments: the stimuli were blocked separately by stimulus type (i.e., speech and instruments respectively). Each stimulus block was presented 3 times with anger, happiness and sadness being the target emotion respectively. That is, in each stimulus block, stimuli of all 4 emotions were presented randomly, with anger as the target emotion for the first time of presentation, happiness as the target emotion for the second time of presentation, sadness as the target emotion for the third time of presentation. In each time of presentation, each target emotion had 4 representations (i.e., the top 4 rated instruments or speech in each emotion, according to the behavioural experiment). Each representation of each emotion was presented 20 times. Altogether there were 3 (target emotions) * 4 (representations of each target emotion) * 20 (repetitions of each representation of each emotion) * 3 (repetitions of each stimulus block) * 2 (stimulus blocks: speech, instruments) = 1440 trials. The stimuli were presented randomly with the constraint that the same stimulus was not presented consecutively and at least there were two different stimuli between identical stimuli. The Inter-Stimulus-Interval (ISI) was 1000 ms. The participants were given a go/no go task, i.e., they monitored for the target emotion in each presentation session by pressing a button as quickly as possible. Prior to recording, participants had a two-minute practice session to familiarize themselves with the task.

Experiment 2

The second experiment aimed to more directly compare emotional speech and musical instruments with the priming paradigm: musical instrumental sounds were used as primes and emotional speech as targets, following a similar approach where words were used targets and musical sounds were primes (Steinbeis and Koelsch, 2011). Moreover, in this study, we used an explicit priming paradigm, i.e., tasks that directly require participants to judge the relatedness between primes and targets. The reason is that the N400 effect could be either absent (Painter and Koelsch, 2011) or small (Frey et al., 2014) if an implicit priming paradigm (i.e., tasks unrelated to the judgement of the relatedness between primes and targets) was used. There were altogether 9 instrument-speech pairs (AA=angry instrument-angry speech; HA=happy instrument-angry speech; SA=sad instrument-angry speech; AH=angry instrument-happy speech; HH=happy instrument-happy speech; SH=sad instrument-happy speech; AS=angry instrument-sad speech; HS=happy instrument-sad speech; SS=sad instrument-sad speech). In each pair, each emotion had 4 representations (i.e., the top 4 rated instruments or speech in each emotion, according to the behavioural experiment). Each representation of each pair was presented 20 times. Altogether there were 9 (pairs) * 4 (speech) * 4 (instruments) * 20 (repetitions) = 2880 trials. They were grouped pseudorandomly and presented in four blocks. The Inter-Stimulus-Interval between the prime and target was 1000ms. After hearing each instrument-speech pair, the participants had 1000ms to judge whether the emotions conveyed by the instrument and speech were congruent or not by pressing a button (left = yes, right = no). Prior to recording, participants (different from those for experiment 1) had a two-minute practice session to familiarize themselves with the task.

EEG recording and data analyses for Experiments 1and 2

The EEG was recorded using a Biosemi ActiveTwo system with 64 Ag-AgCI electrodes mounted on an elastic cap. The offsets at each electrode were kept between +/-20 mV. To detect eye movement-related artifacts, bipolar horizontal and vertical EOGs (electro-oculograms) were recorded. The average of left and right mastoids was used as the off-line reference to all electrodes. Analysis software was EEGLAB v. 12.0.2.04b (Delorme and Makeig, 2004). The data was filtered off-line by a band-pass filter of 0.5-30 Hz. Trials with EOG-artifacts were rejected offline using the artifact detection tools in ERPLAB v. 3.0.2.1 (Lopez-Calderon and Luck, 2014). The moving window peak-to-peak threshold tool (moving window width: 200 ms, voltage threshold: 100 μV, window step: 20 ms) and the step-like artifacts tool (moving window width: 400 ms, voltage threshold: 35 μV, window step: 10 ms) were used to reject trials with these artifacts. On average 19% of the data was rejected for anger, 16% was rejected for happiness and 18% was rejected for sadness. ERPs were averaged from the time window of 200 ms pre-stimulus onset to 800 ms post-stimulus onset. The EEG epochs were time-locked to the stimulus onset, and baseline corrected (-200 to 0 ms).

There were nine regions of interests (ROIs) for this study: left frontal (LF) electrode-sites (F7, F5, F3, FT7, FC5, FC3); left central (LC) electrode-sites (C5, C3, TP7, CP5, CP3); left posterior (LP) electrode-sites (P7, P5, P3, PO7, PO3); right frontal (RF) electrode-sites (F4, F6, F8, FC4, FC6, FT8); right central (RC) electrode-sites (C4, C6, CP4, CP6, TP8); right posterior (RP) electrode-sites (P4, P6, P8, PO4, PO8); midline frontal (MF) electrode-sites (F1, Fz, F2, FC1, FCZ, FC2); midline central (MC) electrode-sites (C1, Cz, C2, CP1, CPZ, CP2); midline posterior (MP)

electrode-sites (P1, PZ, P2, POZ). The ERP data was averaged according to the nine ROIs for analyses of variance (ANOVAs).

3.3.2 Results of experiment 1: P200 and LPC

Behavioural results

The mean error rate for angry speech was 2.9% (SD=0.5), happy speech was 3.6% (SD=0.4), sad speech was 3.8% (SD=0.8), angry instruments was 14.7% (SD=1), happy instruments was 15.1% (SD=1.5), and sad instruments was 15.9% (SD=1.2). A two-way ANOVA (stimulus type and emotion) for the error rate showed that the main effects of stimulus type and emotion were significant (type: $F(1, 14) = 2367.34$, $p<0.001$; emotion: $F(2, 28) = 9.61$, $p<0.01$) while the interaction between them was non-significant ($F(2, 28) = 0.56$, $p=0.58$). Average reaction time for angry speech was 808ms (SD=5.5), happy speech was 811ms (SD=4), sad speech was 805ms (SD=4.9), angry instruments was 813ms (SD=4.4), happy instruments was 804ms (SD=3.8), and sad instruments was 810ms (SD=4.6). A two-way ANOVA (stimulus type and emotion) for reaction time showed that neither of the two factors was significant (stimulus type: $F(1, 14) = 1.04$, $p=0.33$; emotion: ($F(2, 28) = 3.23$, $p=0.06$).

P200

P200 was quantified using a mean amplitude from 170 to 230 ms from stimulus onset. The selection of the interval was based on visual inspection and previous reports (Paulmann et al., 2013). The ERP waveforms, scalp topography and the mean amplitude are displayed in Figures 3.2-3.4.

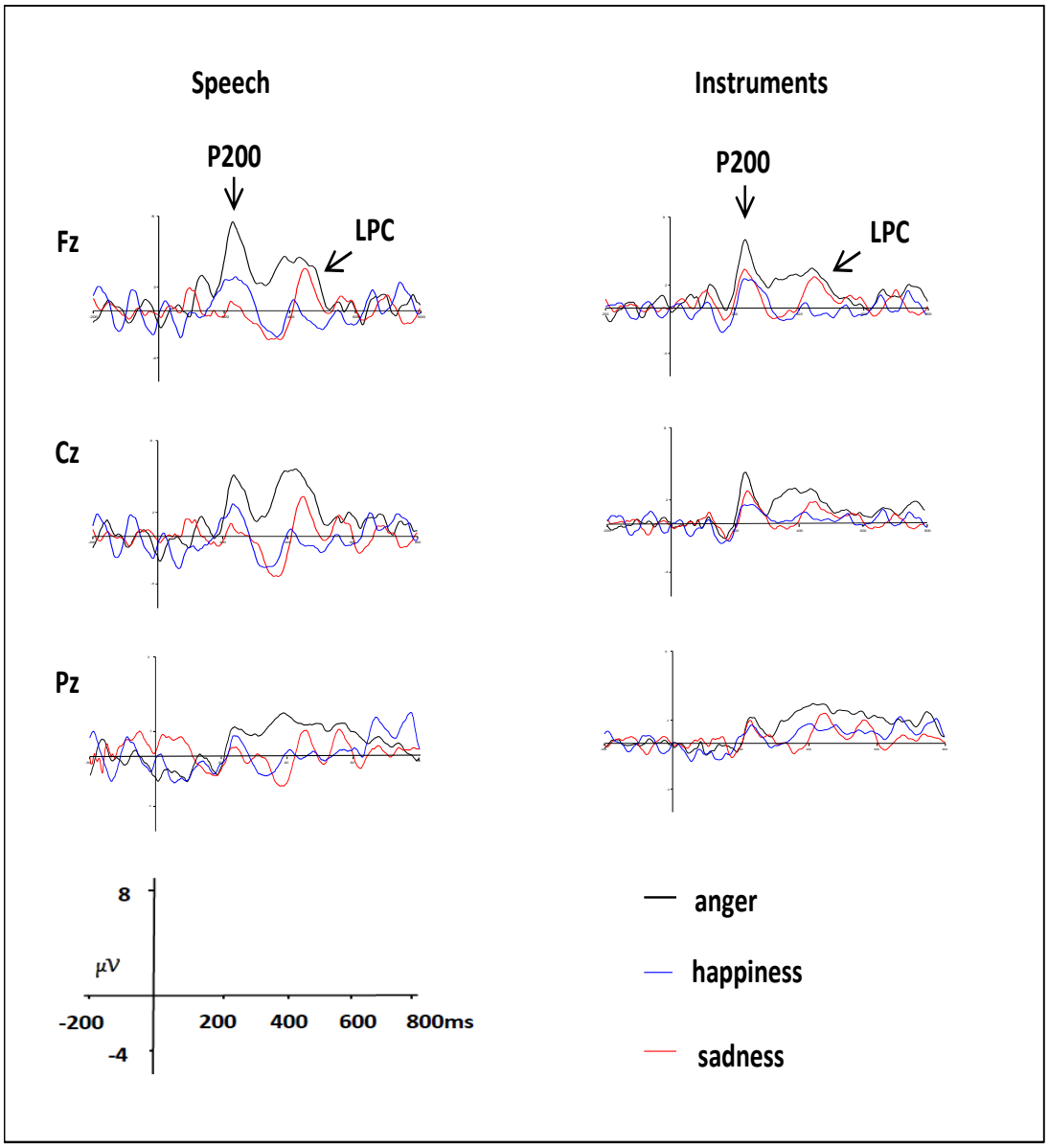


Figure 3.2 ERP waveforms demonstrating the main effects of P200 and LPC at selected electrodes for speech and instruments under the conditions of anger, happiness and sadness.

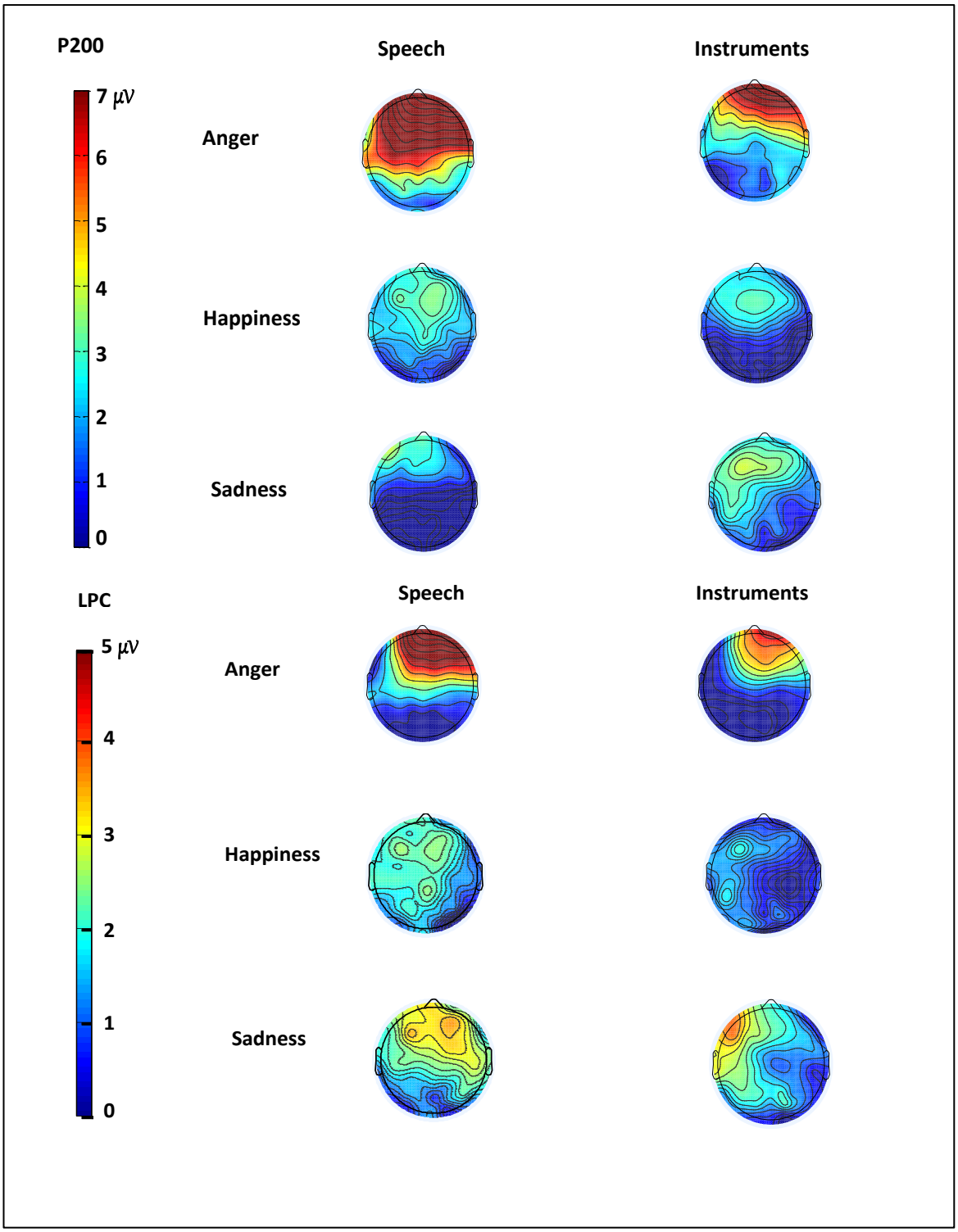


Figure 3.3 Scalp topography of the P200 and LPC for speech and instruments under the conditions of anger, happiness and sadness.

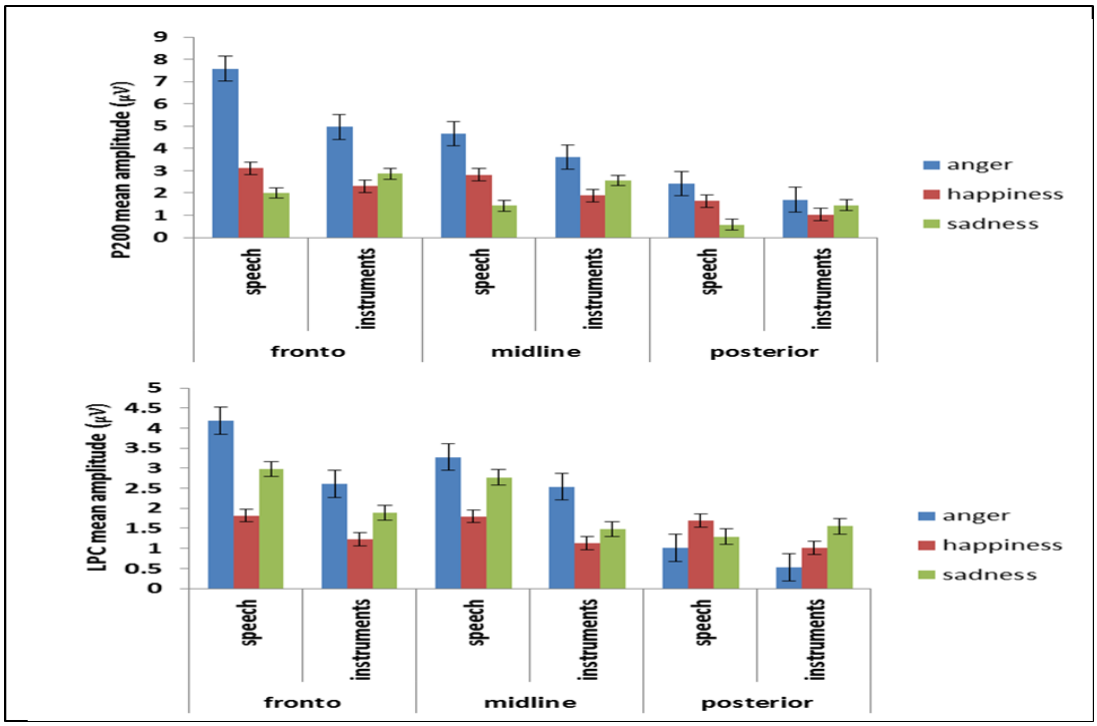


Figure 3.4 The mean amplitude of P200 and LPC for speech and instruments under the conditions of anger, happiness and sadness. Error bars represent the standard error of the mean.

A three-way (type: speech and instruments; emotion: anger, happiness and sadness; and ROI: LF, MF, RF, LC, MC, RC, LP, MP, RP) repeated measures ANOVA was performed on the peak latency and mean amplitude of P200. For peak latency, no significant main effects (emotion, type and ROI) were found. However, for mean amplitude, significant main effects were found: type ($F_{(1, 14)} = 15.1$, $p < 0.01$, $\eta_p^2 = 0.52$), emotion ($F_{(2, 28)} = 334.6$, $p < 0.001$, $\eta_p^2 = 0.96$), ROIs ($F_{(8, 112)} = 153.64$, $p < 0.001$, $\eta_p^2 = 0.92$). Their interaction type * emotion * ROIs was also significant ($F_{(16, 224)} = 10.4$, $p < 0.001$, $\eta_p^2 = 0.43$).

Post-hoc tests showed more details: with regard to type, speech had significantly higher P200 amplitude than instruments in anger ($F_{(1, 14)} = 64.11$, $p < 0.001$, $\eta_p^2 =$

0.82) and happiness ($F_{(1, 14)} = 15.31$, $p < 0.01$, $\eta_p^2 = 0.52$). The opposite was true in sadness ($F_{(1, 14)} = 11.27$, $p < 0.01$, $\eta_p^2 = 0.45$). In terms of emotion, speech and instruments had similar patterns with regard to anger: angry speech had significantly higher P200 amplitude than happy ($F_{(1, 14)} = 125.1$, $p < 0.001$, $\eta_p^2 = 0.9$) and sad ($F_{(1, 14)} = 464.3$, $p < 0.001$, $\eta_p^2 = 0.97$) speech; angry instruments were significantly higher than happy ($F_{(1, 14)} = 133.8$, $p < 0.001$, $\eta_p^2 = 0.91$) and sad ($F_{(1, 14)} = 64.66$, $p < 0.001$, $\eta_p^2 = 0.82$) instruments. Happy speech had significantly larger P200 amplitude than sad speech ($F_{(1, 14)} = 13.37$, $p < 0.01$, $\eta_p^2 = 0.49$) while in the instrument condition, the opposite was true: happy instrument had significantly lower P200 amplitude than sad instruments ($F_{(1, 14)} = 17.93$, $p < 0.01$, $\eta_p^2 = 0.56$). Table 3.3 shows the results of post-hoc tests at each ROI. It can be observed that for speech, the overall pattern was A>H>S across all the ROIs. The differences between each emotion and another were particularly prominent (i.e., significant) at fronto-central areas. For musical instruments, the overall pattern was A>S>H across all the ROIs. Similarly to speech, the differences were more pronounced (i.e., significant) at fronto-central ROIs.

Table 3.3 The results of post-hoc tests at each ROI in terms of the P200 amplitude (A=anger, H=happiness, S=sadness, statistically significant comparisons are in bold, $p<0.017$).

P200		LF	RF	LC	RC	LP	RP	MF	MC	MP
Speech	A vs. H	**A>H**	**A>H**	**A>H**	**A>H**	A>H	**A>H**	**A>H**	**A>H**	**A>H**
	A vs. S	**A>S**	**A>S**	**A>S**	**A>S**	**A>S**	**A>S**	**A>S**	**A>S**	**A>S**
	H vs. S	**H>S**	**H>S**	H>S	H>S	H>S	H>S	**H>S**	**H>S**	H>S
Instruments	A vs. H	**A>H**	**A>H**	**A>H**	**A>H**	A>H	A>H	**A>H**	**A>H**	A>H
	A vs. S	**A>S**	**A>S**	A>S	A>S	A>S	A>S	**A>S**	**A>S**	A>S
	H vs. S	S>H	S>H	S>H	**S>H**	**S>H**	**S>H**	S>H	S>H	**S>H**

LPC

LPC (late positive complex) was measured from 450-750ms after stimulus onset based on visual inspection and previous reports (Paulmann et al., 2013). The ERP waveforms, the mean amplitude and scalp topography are displayed in Figures 3.2-3.4. A three-way (type, emotion and location) repeated measures ANOVA was performed on the peak latency and mean amplitude of LPC. The results were similar to those of P200: for peak latency, no significant main effects (emotion, type and location) were found. However, for mean amplitude, significant main effects were found: type ($F_{(1, 14)} = 35.77$, $p < 0.001$, $\eta_p^2 = 0.94$), emotion ($F_{(2, 28)} = 143.38$, $p < 0.001$, $\eta_p^2 = 0.91$), ROIs ($F_{(8, 112)} = 128.71$, $p < 0.001$, $\eta_p^2 = 0.9$). The interaction type * emotion * ROIs was also significant ($F_{(16, 224)} = 3.78$, $p < 0.001$, $\eta_p^2 = 0.21$) (see Figure 3.2 for the ERP waveforms and Figure 3.3 for the scalp topography).

Post-hoc contrasts show that in all emotion conditions, speech had significantly higher LPC amplitude than music (anger: $F_{(1, 14)} = 87.99$, $p < 0.001$, $\eta_p^2 = 0.86$; happiness: $F_{(1, 14)} = 68.31$, $p < 0.001$, $\eta_p^2 = 0.83$; sadness: $F_{(1, 14)} = 73.57$, $p < 0.001$, $\eta_p^2 = 0.84$). With regard to emotion, speech and instruments present a similar picture: anger was significantly higher than happiness (speech: $F_{(1, 14)} = 119.9$, $p < 0.001$, $\eta_p^2 = 0.9$; instruments: $F_{(1, 14)} = 133.99$, $p < 0.001$, $\eta_p^2 = 0.91$) and sadness (speech: $F_{(1, 14)} = 35.42$, $p < 0.001$, $\eta_p^2 = 0.72$; instruments: $F_{(1, 14)} = 52.09$, $p < 0.001$, $\eta_p^2 = 0.79$); sadness was significantly higher than happiness (speech: $F_{(1, 14)} = 34.73$, $p < 0.001$, $\eta_p^2 = 0.71$; instruments: $F_{(1, 14)} = 53.11$, $p < 0.001$, $\eta_p^2 = 0.79$). Table 3.4 shows the results of post-hoc tests at each ROI. It can be observed that for both speech and musical instruments, the pattern of A>S>H was present across all fronto-central ROIs.

Table 3.4 The results of post-hoc tests at each ROI in terms of the LPC amplitude (A=anger, H=happiness, S=sadness, statistically significant comparisons are in bold, $p < 0.017$).

LPC		LF	RF	LC	RC	LP	RP	MF	MC	MP
Speech	A vs. H	**A>H**	**A>H**	**A>H**	**A>H**	A<H	A<H	**A>H**	**A>H**	A<H
	A vs. S	**A>S**	**A>S**	**A>S**	**A>S**	A<S	A<S	**A>S**	**A>S**	A<S
	H vs. S	S>H	**S>H**	S>H	S>H	S>H	**S>H**	**S>H**	**S>H**	S>H
Instruments	A vs. H	**A>H**	**A>H**	**A>H**	**A>H**	A<H,	A<H	**A>H**	**A>H**	A<H
	A vs. S	**A>S**	**A>S**	A>S	**A>S**	A<S	A<S	**A>S**	**A>S**	A<S
	H vs. S	S>H	**S>H**	S>H	**S>H**	**S>H**	**S>H**	**S>H**	S>H	**S>H**

3.3.3 Results of the ERP experiment 2: N400

Behavioural results

The mean error rate for AA (angry instrument-angry speech) pair was 3.1% (SD=0.6); AH (angry instrument-happy speech) pair was 3.9% (SD=0.6); AS (angry instrument-sad speech) pair was 2.9% (SD=1.1); HA (happy instrument-angry speech) pair was 5.7% (SD=0.9); HH (happy instrument-happy speech) pair was 4.1% (SD=0.8); HS (happy instrument-sad speech) pair was 3.8%; (SD=0.6); SA (sad instrument-angry speech) pair was 8.2% (SD=1.1); SH (sad instrument-happy speech) pair was 7.8% (SD=0.9); SS (sad instrument-sad speech) pair was 8.1% (SD=1.1). A two-way (prime and target) repeated measures ANOVA showed that the effects of prime, target and their interaction were significant: prime ($F(2, 28) = 292.2$, $p < 0.001$); target ($F(2, 28) = 5.93$, $p < 0.01$); interaction ($F(4, 56) = 8.46$, $p < 0.001$). This could suggest that the listeners' judgement accuracy on the

congruence/incongruence of the target depends on the prime. No reaction time data was collected because the task was a delayed response task.

N400

The selection of N400 time window in this study was from 350 to 500ms based on visual inspection and previous literature on music and language priming (Painter and Koelsch, 2011). The N400 appeared larger in amplitude for emotionally incongruous instrument-speech pairs than the congruous pairs (see Figure 3.5 for the ERP waveforms, Figure 3.6 for the mean amplitude and Figure 3.7 for the scalp topography).

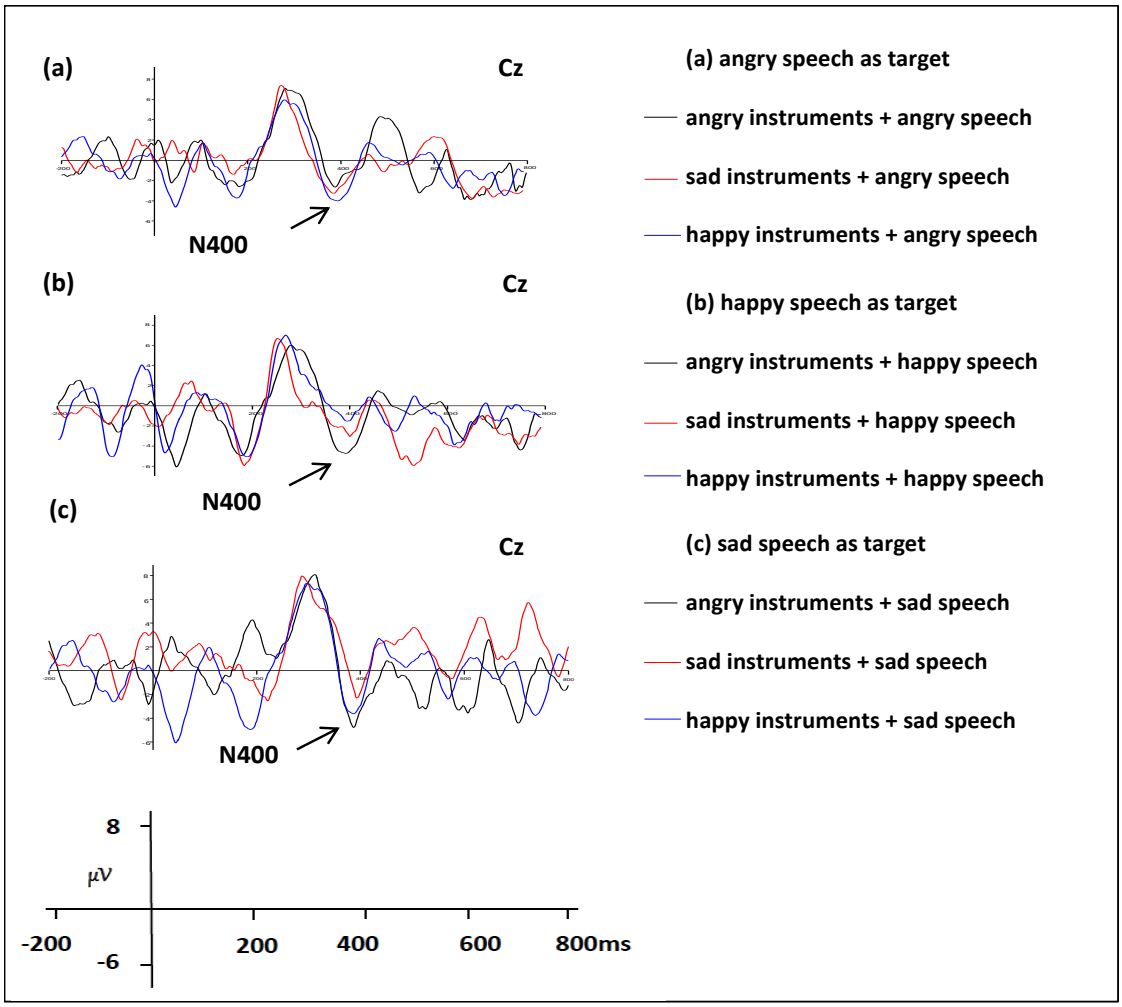


Figure 3.5 The N400 effect at Cz when angry speech (a), happy speech (b) and sad speech (c) was the target primed by musical instruments of different emotional categories.

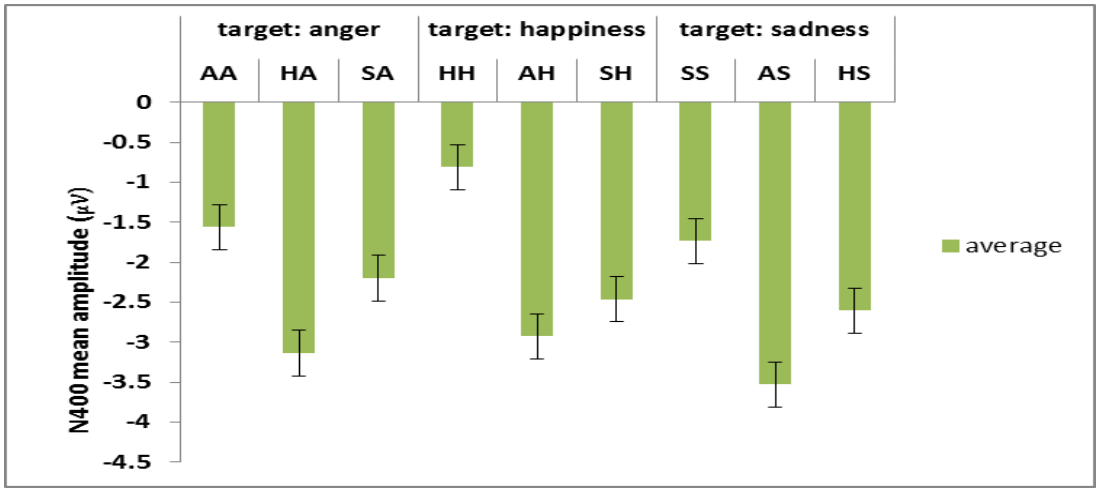


Figure 3.6 The mean amplitude of the N400 for the nine instrument-speech pairs (AA=angry instrument-angry speech; AH=angry instrument-happy speech; AS=angry instrument-sad speech; HA=happy instrument-angry speech; HH=happy instrument-happy speech; HS=happy instrument-sad speech; SA=sad instrument-angry speech; SH=sad instrument-happy speech; SS=sad instrument-sad speech). Error bars represent the standard error of the mean.

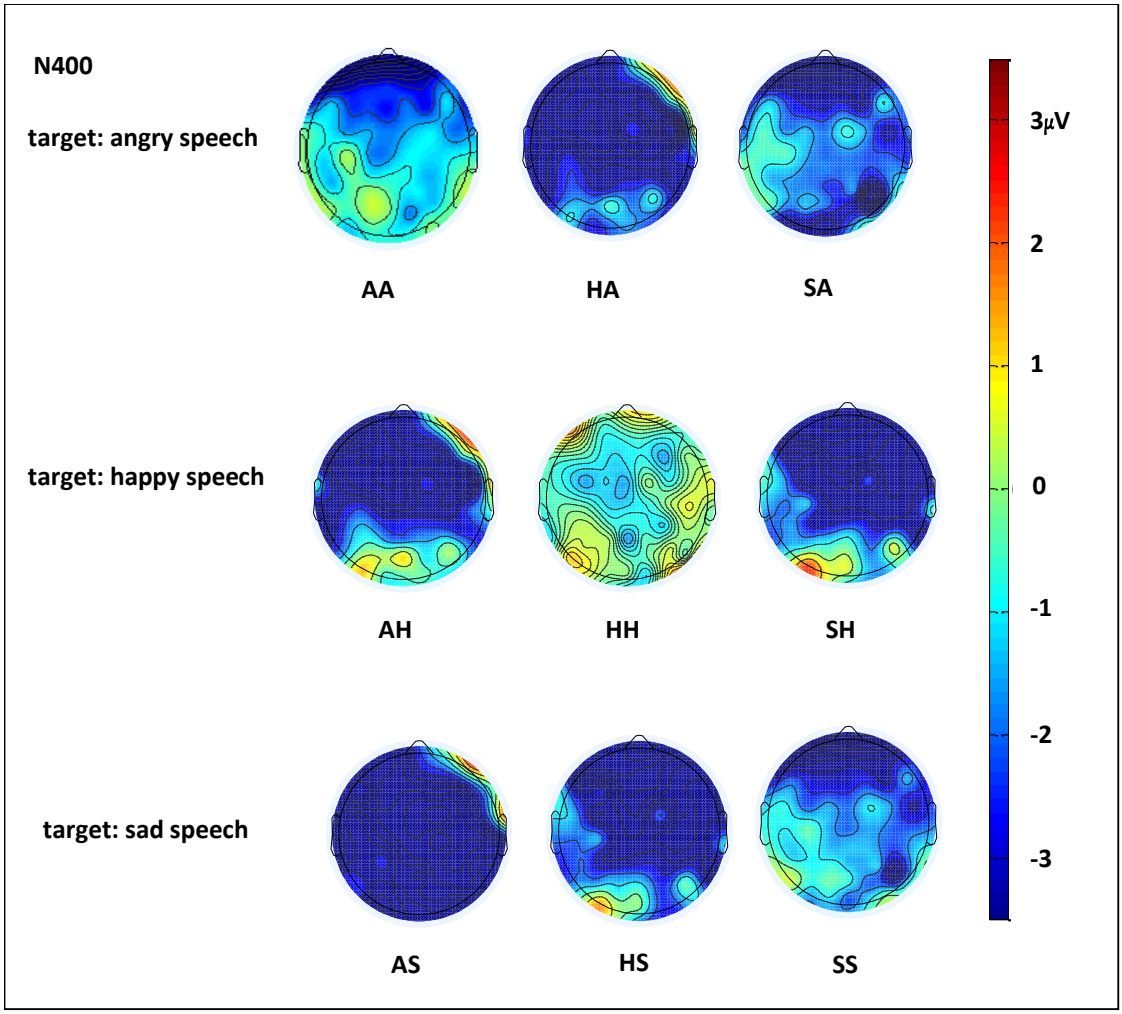


Figure 3.7 Scalp topography of the N400 effect for the nine instrument-speech pairs (AA=angry instrument-angry speech; AH=angry instrument-happy speech; AS=angry instrument-sad speech; HA=happy instrument-angry speech; HH=happy instrument-happy speech; HS=happy instrument-sad speech; SA=sad instrument-angry speech; SH=sad instrument-happy speech; SS=sad instrument-sad speech).

A three-way (target, prime and ROIs) repeated measures ANOVA was carried out on the mean amplitude of N400 between 350ms and 500ms. The effects of target, prime and ROIs were significant: target ($F_{(2, 28)} = 205.22$, $p < 0.001$, $\eta_p^2 = 0.94$); prime ($F_{(2, 28)} = 166.76$, $p < 0.001$, $\eta_p^2 = 0.92$); ROI ($F_{(8, 112)} = 148.94$, $p < 0.001$, $\eta_p^2 = 0.91$). The interaction between the prime and target was significant ($F_{(4, 56)} = 604.75$, $p < 0.001$, η_p^2

= 0.98), suggesting that the modulation of the N400 amplitude of the target depended on the prime.

Post-hoc contrasts revealed more details of the N400 amplitude comparisons between different instrument-speech pairs (AA=angry instrument-angry speech; AH=angry instrument-happy speech; AS=angry instrument-sad speech; HA=happy instrument-angry speech; HH=happy instrument-happy speech; HS=happy instrument-sad speech; SA=sad instrument-angry speech; SH=sad instrument-happy speech; SS=sad instrument-sad speech): AA vs. HA ($F_{(1, 14)} = 364.67$, $p < 0.001$, $\eta_p^2 = 0.96$); AA vs. SA ($F_{(1, 14)} = 59.16$, $p < 0.001$, $\eta_p^2 = 0.81$); SA vs. HA ($F_{(1, 14)} = 122.21$, $p < 0.001$, $\eta_p^2 = 0.9$); AH vs. HH ($F_{(1, 14)} = 666.39$, $p < 0.001$, $\eta_p^2 = 0.99$); AH vs. SH ($F_{(1, 14)} = 56.82$, $p < 0.001$, $\eta_p^2 = 0.8$); HH vs. SH ($F_{(1, 14)} = 558.23$, $p < 0.001$, $\eta_p^2 = 0.98$); AS vs. HS ($F_{(1, 14)} = 426.41$, $p < 0.001$, $\eta_p^2 = 0.97$); AS vs. SS ($F_{(1, 14)} = 481.81$, $p < 0.001$, $\eta_p^2 = 0.98$); HS vs. SS ($F_{(1, 14)} = 452.52$, $p < 0.001$, $\eta_p^2 = 0.97$).

3.3.4 Summary of the results of the ERP experiment 1 and 2

The results of the ERP experiment 1 showed that firstly, emotional speech triggered larger P200 and LPC amplitudes than musical instruments, with the exception that sad instruments triggered larger P200 than sad speech. Secondly, the results on P200 and LPC showed that for both speech and instrument conditions, anger overall had significantly higher P200 and LPC amplitude than happiness and sadness. Moreover, for the P200, happiness was significantly higher than sadness in the speech condition while the opposite was true for the musical instrument condition; for the LPC, happiness was significantly lower than sadness in both speech and musical instrument conditions. The results of the experiment 2 demonstrated that emotionally

congruous instrument-speech pairs triggered smaller N400 amplitude than emotionally incongruous instrument-speech pairs.

3.4 Discussion and conclusion

3.4.1 The behavioral experiment

In this study we aimed to explore emotional connotations of musical timbre of isolated instrument sounds through the perspective of affective speech using behavioral and ERP experiments. The results of the behavioral experiment suggested that there were no significant differences between affective speech and instrument sounds with regard to the six timbral features in each emotion category (i.e., anger, happiness and sadness). Moreover, the patterns of emotions were similar for speech and instruments across all of the six timbral features, especially with regard to anger and happiness (Table 3.2). The results showed that angry speech/instrument sounds had higher attack slope and high-low frequency energy ratio than happy speech/instrument sounds. This is in line with the report on affective speech (Banse and Scherer, 1996; Goudbeek and Scherer, 2010; Laukka et al., 2005; Scherer, 1989) that the more activated an emotion is (e.g., anger), the higher the attack slope and high-low frequency energy ratio, probably due to the high physiological arousal triggered by more activated emotions such as anger (Banse and Scherer, 1996; Scherer, 1989). Moreover, the results on speech are also consistent with the report that (Goudbeek and Scherer, 2010; Laukka et al., 2005; Xu et al., 2013b) the valence of emotion is often negatively correlated with high-low frequency energy ratio, i.e., the more positive an emotion is (e.g., happiness), the lower the ratio. This pattern also holds true for musical timbre: sound spectrums with greater amount of high frequency energy are generally perceived as sharp and angry (Juslin, 1997). In

addition, musical sounds with positive valence were found to have lower attack slope and lower high-low frequency energy ratio (Eerola, et al., 2012). With regard to sadness, previous reports show that sad speech and music are often characterized by legato articulation (Bresin and Friberg, 2000; Juslin, 1997, 2000), i.e., lower degree of attack slope. This is supported by this study in terms of sad speech but not sad instruments. In addition, this study showed that sadness had the lowest high-low frequency energy ratio for speech but not for musical instruments, which does not agree with the report that in music, low amount of high frequency leads to the perception of soft timbre and sadness (Juslin, 1997). The reason can be attributed to the fact that there are usually two kinds of sadness: depressed sadness and grieving sadness (Scherer, 1979; Xu et al., 2013a). In this study, sad speech was produced more like depressed sadness while sad instruments rated by the listeners sounded more like grieving sadness (i.e., with more energy). Depressed sadness is usually characterized by low amount of high spectral energy while the opposite is true for grieving sadness (Scherer, 1979), which is consistent with the findings of this study. It is worth noting that anger and happiness also have different modes of expressions (i.e., hot vs. cold anger; pleasure vs. joy). In this study, the speakers produced hot anger and joy which were in the same acoustic direction as musical instruments selected for anger and happiness respectively. Hence, there were more consistencies between speech and music in terms of anger and happiness than sadness.

In terms of spectral skewness and centroid of spectral gravity, the results on speech and instruments shared the same direction: for spectral centroid, anger had the highest value while sadness the lowest; for skewness, sadness had the highest value while anger the lowest. Happiness was in between for both measures. The results on speech are in line with previous reports (Xu et al., 2013b) that happiness is

associated with higher degree of spectral tilt (skewness) and lower spectral centroid than anger. Moreover, greater spectral skewness suggests breathiness in voice (Xu et al., 2013b). Musical instruments perceived as conveying sadness in this study included the flute, oboe and bassoon, whose acoustic characteristics can be very similar to breathy voice. Musical instruments perceived as expressing anger and happiness included brass, keyboard and percussion instruments, whose degree of breathiness is much lower than the instruments conveying sadness. This could explain why in this study, sad instruments had the highest degree of spectral skewness. The results on musical instruments, however, were in opposite direction from the findings in Eerola et al. (2012) where positively valenced musical sounds (i.e., pleasant sounds) were associated with higher spectral centroid than unpleasant sounds. The reasons could be that in their study, a dimensional approach to emotion was adopted (Eerola et al., 2012, p.51) rather than a discrete approach as in this study. Therefore, what was perceived as positively valenced sounds in their study could fall into more discrete categories than happiness alone in this study. The larger data set of their study could lead to inconsistency with the results of this study. The results on spectral flux and spectral regularity showed relatively consistent patterns for speech and instruments: happiness was associated with higher degree of spectral flux and lower spectral regularity than anger and sadness. This is consistent with previous reports on affective speech (Xu et al., 2013b) and musical instruments (Eerola et al., 2012) where positively valenced sounds are spectrally more dynamic (i.e., more fluctuating and less regular).

3.4.2 ERP experiments

P200

This study firstly showed that different vocal emotions and musical timbre can be differentiated by the P200 amplitude. In particular, both angry speech and angry instrument sounds had the highest P200 amplitude compared to happy and sad speech/instrument sounds. This is consistent with previous reports on affective speech where highly arousing emotions such as anger often trigger significantly higher P200 amplitude than less arousing emotions such as sadness (Paulmann, et al., 2013; Sauter and Eimer, 2010; Spreckelmeyer et al., 2006). In terms of music, highly aroused emotion such as anger is often associated with faster speed, higher intensity and greater roughness than less aroused emotions in a way similar to speech (Juslin and Laukka, 2003; Juslin and Västfjäll, 2008). It is therefore not surprising that the P200 amplitude of angry instrument sounds triggered the highest amplitude in a way similar to angry speech, as revealed in this study. With regard to happiness and sadness, happy instrument sounds had lower P200 amplitude than sad instrument sounds while the opposite pattern was true for speech. As discussed in section 3.4.1, the reasons can be attributed to the two kinds of sadness portrayed by instruments and speech, respectively. Sad instruments convey grieving sadness while sad speech conveys depressed sadness, as evidenced from the results of the behavioral study reported above. Therefore, it is reasonable that the ERP results revealed correspondingly different (opposite) P200 patterns for music and speech in terms of sadness.

This finding also supports the fact that the P200 is associated with the human brain's rapid detection of and enhanced attention to emotional stimuli (Paulmann et al., 2013;

Schirmer et al., 2013). In terms of speech, previous studies have reported that affective speech prosody can be differentiated from neutral speech prosody as early as 200 ms, usually at the fronto-central location of the brain (Paulmann and Kotz, 2008; Paulmann et al., 2011). Recent findings (Paulmann et al., 2013) have also reported that different emotions can be differentiated from each other within 200 ms. Cognitive processing of emotional visual stimuli follows a similar pattern: highly (positive and negative) arousing pictures trigger larger P200 than lowly (positive and negative) arousing pictures (Feng et al., 2012). With regard to musical timbre, Meyer et al. (2006) found that the P200 was associated with the differentiation of the emotional connotations of musical timbre. Studies have also reported that 200ms-long musical excerpts with intact temporal and high frequency spectral information were enough to enable listeners to recognize the titles of popular songs (Schellenberg et al., 1999). Such fast cognitive processing of musical timbre can be attributed to the possibility that compared with pitch, timbre can be a robust cue facilitating music feature perception (Robinson and Patterson, 1995).

There may be questions about whether the P200 reflects detection of physical differences between sound stimuli (i.e., low level acoustic differences of the stimuli) or cognitive interpretation of the emotional connotation of the sound stimuli. Theoretical (Schirmer and Kotz, 2006) and experimental (Paulmann and Kotz, 2008; Pulvermüller and Shtyrov, 2006) studies on affective speech perception has suggested that early ERP components such as P200 reflects the processing of both physical properties of the sound stimuli and higher-order cognitive events such as emotion, due to the rapid differentiation of emotional sentences from neutral sentences within the time window as early as 200ms. Recent research using Mandarin sentences (Jiang et al., 2014) separately tested the roles of physical

properties and emotional categories in perception of affective speech stimuli. The results demonstrate that both physical and emotional aspects of the sound stimuli can be detected at an early stage (i.e., around 200 ms). With regard to music, the P200 was found to reflect the emotional connotations inherent of musical timbre (Meyer et al., 2006) due to the finding that additional anterior cingulate cortex (ACC) activities in the P200 time window were observed in the perception of musical timbre (Meyer et al., 2006). ACC activities have been particularly associated with cognitive processing of emotion related stimuli (Phan et al., 2002). Therefore, it is plausible that musical timbre perception not only activates auditory areas for processing the physical properties of the sound stimuli but also triggers response from areas responsible for processing higher cognitive dimensions such as emotion (Meyer et al., 2006).

LPC

The results of this study showed that for both speech and musical instrument sounds, anger had the highest LPC amplitude while happiness the lowest, with sadness in between. The results are consistent with previous studies showing that highly arousing emotions such as anger usually trigger larger LPC amplitude than lowly arousing emotions such as sadness (Hinojosa et al., 2009; Paulmann et al., 2013; Rozenkrants et al., 2008). More recent findings on affective speech and vocalizations also suggest anger tends to trigger larger LPC than other emotions (Pell et al., 2015). Nevertheless, the finding of this study that sadness had larger LPC amplitude than happiness is the opposite of previous findings (e.g., Paulmann et al., 2013). For musical instruments, the explanation for this could be attributed to the fact the sad instruments selected for this study conveyed grieving sadness which could sound much more activated and aroused than happiness (as discussed in section 3.4.1),

leading to the corresponding ERP patterns where sad instruments triggered larger LPC than happy instruments. For speech, the explanation requires more discussion about the function of LPC. LPC is proposed to be a part of multi-step model of emotion processing (Schirmer and Kotz, 2006). If the P200 is a reflection of the processing of both physical properties and emotional connotations of the stimuli as discussed above, then the LPC is more indicative of the enhanced and continuous cognitive evaluation of emotional stimuli (Jiang et al., 2014; Kotz and Paulmann, 2011; Paulmann et al., 2013; Schirmer and Kotz, 2006). Specifically, the P200 is more about the detection of the emotional salience of the sound stimuli while the LPC is responsible for the fine-tuned analysis of the emotional meaning conveyed thereafter in order to ensure appropriate actions (Kotz and Paulmann, 2011; Paulmann et al., 2013). The more enhanced LPC of sad speech than happy speech as seen in this study could reflect a more pronounced cognitive evaluation of the speaker's sad speech (depressed sadness), possibly due to its potential implication to beg for sympathy and require subsequent consolation from the listener (Xu et al., 2013a). Admittedly, this pattern does not necessarily exist universally and could vary greatly from individual to individual, and hence there could well be experimental inconsistencies between studies in this regard.

N400

The results of the second ERP experiment showed a clear N400 effect elicited by affective incongruence between affective speech and musical instrument sounds. The effect was particularly pronounced for speech-instrument pairs with more opposing emotions respectively: the N400 amplitude of angry-sad and happy-sad pairs was significantly larger than that of other pairs. The reason could be that anger and happiness share many similar acoustic features since both of them are emotions with

a high level of activity, and hence perceptually they can be sometimes indistinguishable (cf. Scherer, 2003). Therefore, the acoustic contrast between anger and happiness could be less than that between anger and sadness or happiness and sadness. The N400 patterns, correspondingly, reflected such differences.

The finding is also in the same direction as that of Steinbeis and Koelsch (2011) where affective priming paradigm was used to show that musical instrument timbre could communicate emotion to musically trained and untrained listeners. Words (visually presented) that were emotionally incongruent with the timbre of musical chords generated larger N400 amplitude than emotionally congruent words-music pairs. Similar findings were reported in Painter and Koelsch (2011). Nevertheless, the music stimuli in the aforementioned studies did not have a strict control for other acoustic features such as fundamental frequency, intensity or duration. This means it could be the cohort of all acoustic features (not just timbre) that contributed to the N400 effect. The present study, in contrast, strictly controls all acoustic features except timbre, thus demonstrating a clearer picture of the emotional connotations of musical instrument timbre. In addition, this study further extends previous affective priming research on emotional meanings of music by showing that when targets were auditorily presented affective speech (rather than visually presented words), the N400 could still be elicited due to the emotional incongruence between music and speech. Hence, the results provide further evidence that the N400 effect can exist regardless of domain or modality differences (Cummings et al., 2006; McPherson and Holcomb, 1999; Painter and Koelsch, 2011).

3.4.3 The processing advantage of human voice

This study shows that on the whole, affective speech triggered larger P200 and LPC amplitudes than musical instruments. This finding is in line with previous studies reporting the cognitive processing advantage of human voice compared with other types of sound stimuli. Behavioral experiments on reaction time differences between instrumental sounds (e.g., strings, percussion) and vocal sounds (sung vowels) show that voice stimuli elicit faster reaction time than musical instruments (Agus et al., 2012). Evidence from EEG/ERP experiments also supports the brain's processing preference for voice: compared with musical instruments, human voice can elicit a significantly greater positive component peaking at 320 ms (Levy et al., 2001), particularly at the frontal area. Compared with environmental sounds and bird songs, human voice also triggers greater P200 amplitude in the frontal temporal area (Charest et al., 2009). Research on musical expertise differences between musicians and non-musicians also show that voice stimuli are processed faster than music stimuli by all participants, regardless of musical training (Kaganovich et al., 2013).

The processing advantage of voice can be further supported by the fact that humans are by nature voice experts (Latinus and Belin, 2011), since speech plays a crucial role in human communication (Liberman and Mattingly, 1989). Evidence abounds in developmental research on human voice sensitivity from infancy: infants' voice sensitivity to mother's voice develops even before birth (Kisilevsky et al., 2003); five-month-olds can show enhanced fronto-temporal activity for voice stimuli (Rogier et al., 2010); seven-month-olds begin to show similar language processing patterns as adults (Grossmann et al., 2010); one-year-olds can follow adults' voice direction (Rossano et al., 2012), etc.

3.4.4 Evolutionary implications of the present study

Music and speech are two major platforms of communicating emotion (Juslin and Laukka, 2003), sharing similar evolutionary implications (Darwin, 1871; Cross, 2009a). Evolutionarily, emotion is adapted under selection pressure (Darwin, 1872) as a mechanism for interacting with other living organisms (Ekman, 1992). Emotional vocal expressions are likely selected to have the effect of influencing the receiver for the benefit of the signaller (Morton, 1977; Ohala, 1984; Xu et al. 2013a, 2013b). In particular, vocalizations that mimic a big animal would help to scare off the listener because a larger animal stands a better chance of winning a physical confrontation, and vocalizations that mimic a small animal or even an infant would help to attract the listener by showing lack of threat and eliciting parental instinct (Morton, 1977; Ohala, 1984). Due to simple physical laws, the vocalization of a large animal is likely to have low pitch and rough sound quality, whereas that of a small animal is likely to be high-pitched and pure-tone like (Morton, 1977). Here the rough and tone-like voice quality form a continuum in terms of spectral shape: the rougher the voice, the flatter the spectrum due to the abundance of high-frequency energy; and the more pure-tone like the voice, the more negative the spectral tilt due to lack of high-frequency energy (Stevens, 1998).

In this study angry and happy speech showed similar timbral patterns to those of musical instruments in the same emotion category (i.e., anger corresponded to rough sound quality while happiness corresponded to pure-tone like sound quality). The following ERP studies also reflected that the brain's evaluation for the musical instrument timbre was in the same direction as that for affective speech. Therefore, the results further suggest that the timbre of instrumental sounds could imply bio-evolutionary meanings similar to those of affective speech. That is, similar to

affective speech, angry instrument timbre projects a large body size while happy instrument timbre projects a small body size. Examples can be found in orchestral works (e.g., Tchaikovsky's *The Nutcracker*; Prokofiev's *Peter and the Wolf*) where instruments with rough timbre are used to portray angry characters while instruments with tone-like timbre are used to portray happy characters.

With regard to sadness, the two kinds of sadness (grieving sadness and depressed sadness) have different body size projections: grieving sadness should project a large body size due to its demanding nature, while depressed sadness has relatively neutral size projection due to its lack of communicative intention (Xu et al., 2013a). Furthermore, high-low frequency energy ratio is likely to be reduced further in depressed sadness because reduced vocal effort would result in reduction of high-frequency energy (Traunmüller and Eriksson, 2000). The results of this study were in line with these predictions, as sad speech timbre (depressed sadness) showed the lowest ratio of high-low frequency energy (Figure 3.1) as well as different ERP patterns from sad instrument timbre (grieving sadness).

The ERP patterns of different emotions revealed in this study also provide further evidence that cognitive processing of different emotions is unequal (Lindquist et al., 2012; Vaish et al., 2008). A recent study focusing on the comparison between different threat-related stimuli (e.g., anger and fear) shows that sleeping infants have larger MMN (mismatch negativity) for anger than fear (Zhang et al., 2014). The reason can be associated with different evolutionary functions of anger and fear: anger often triggers confrontational/aggressive approach to fight against danger threatening survival; fear, on the other hand, often triggers a "flight" response so as to avoid danger for self-protection (Darwin, 1872/1965). Therefore, the enhanced ERP response to anger revealed in this study (in both speech and music conditions)

may reflect the enhanced arousal preparing individuals to fight under the mechanism of anger.

Taken together, the acoustic and ERP findings of this study suggest that the timbre of simple, isolated musical instrument sounds can convey emotion in a way similar to affective speech. In addition, the timbral features of both instruments and speech in each emotional category are consistent with the prediction of body-size projection theory on emotion. These findings thus add to the growing evidence that music and speech could share a common code in communicating emotion (Juslin and Laukka, 2003) and both of them could have evolutionary implications (Darwin, 1871; Cross, 2009).

Chapter 4

Perception of pitch prominence and expectation in speech and music

4.1 Introduction

Pitch change is an important source of information about our auditory environment, particularly in terms of speech and music. The rising and falling pitch patterns (i.e., melody) common to both speech and music have naturally given rise to the question as to what relations there may be between speech and music melody (Bolinger,1985). Currently, there are two major views regarding the relations between the two domains: one is that speech and music melody processing share common cognitive resources although the surface representations of the two domains differ (Patel, 2008); the other is that the processing of speech and music melody is largely separate (despite some similarities) due to differences in both surface structure and underlying neurophysiological mechanisms (Peretz, 2006, 2012; Zatorre and Baum, 2012). Evidence for each view mainly comes from studies on congenital amusia (cf. Peretz and Hyde, 2003; Patel, 2008), statistics of pitch patterning (Patel et al., 2006) and neuroimaging of normal and brain impaired individuals (cf. Zatorre and Baum, 2012).

The study in this chapter aimed to shed new light on the above two views by exploring the relations between speech and music melody from a different perspective: pitch prominence and expectation. They play a vital role in guiding the perceptual processing of melodic information in speech and music. This is because pitch prominence arises from sound events that stand out from the acoustic environment due to their prosodic salience (Terken and Hermes, 2000). Such prosodic salience usually helps direct listeners' attention to acoustically important

events such as focus in speech or melodic accent in music, thus facilitating listeners' comprehension of speech or music (Parncutt, 2003). With regard to expectation in the context of acoustic communication, it is a cognitive mechanism enabling listeners to anticipate future sound events (Meyer, 1956). It is one of the essential cognitive abilities for humans to adapt and survive because failure to predict and anticipate future events increases the risk of losing control and decreases the possibility of preparing for dangers (Huron, 2006). Violation of expectation, therefore, is likely to give rise to surprise (Reisenzein, 2000; Scherer et al., 2004). In this study, we will specifically concentrate on prosodic focus in speech (with Mandarin as the target language) and music melodic accent, together with expectation patterns (i.e., the degree of surprise) in both speech and music melody. The background and research questions for this study will be provided in more detail in the following sections.

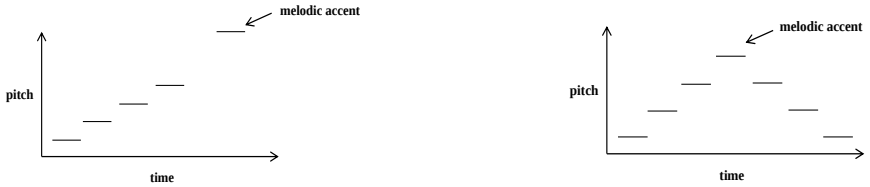
4.1.1 Pitch prominence in speech and music: focus and melodic accent

In speech, focus is an important concept because it serves to highlight the prominence of a piece of information in an utterance, thus facilitating listeners to differentiate the important from the unimportant in the speaker's utterance (Rump and Collier, 1996). One of the essential ways of signalling focus in speech communication is by prosody (Cooper et al., 1985; de Jong, 2004), especially by pitch range expansion as has been evidenced from non-tonal languages (Ladd, 2008; Liberman and Pierrehumbert, 1984) and tonal languages (Chen and Gussenhoven, 2008; Xu, 1999). There has been some evidence for the existence of discrete pitch ranges for functions like focus. For example, Bruce (1977) and Horne (1988) have proposed specific target height of focused components for the sake of speech synthesis. Empirical studies have also provided psychological evidence. For instance,

Rump and Collier (1996) have found that Dutch listeners tended to assign specific pitch values (ranging from 2 to 6 semitones higher than baseline) to focused syllables. ‘t Hart (1981) has found that differences of less than 3 semitones are not significant for the detection of large pitch movement in Dutch. Rietveld and Gussenhoven (1985) have found a smaller threshold, i.e., a pitch difference of 1.5 semitones was sufficient to enable listeners to perceive a difference in Dutch pitch prominence. On the other hand, evidence also exists as to the lack of discriminatory threshold for focus or accent. For example, Ladd and Morton (1997) have found no discriminatory boundary (i.e., threshold) between emphatic and non-emphatic accents in English. There have also been findings of lack of division of pitch range for different types of focus for Dutch (Hanssen, et al., 2008) and English (Sityaev and House, 2003). The above interesting albeit somewhat controversial findings on the threshold of pitch prominence perception in non-tonal languages raises the question as to whether the same pattern could be found in tonal languages such as Mandarin Chinese. So far no empirical research has formally investigated this issue. Given the functional use of F_0 for differentiating lexical words in Mandarin, it could be hypothesized that Mandarin listeners do not necessarily follow the same pitch pattern in pitch prominence perception (e.g., focus) as do listeners of non-tonal languages.

In terms of music, accent is the counterpart of focus. This is because similar to focus in speech, accent in music serves to highlight noticeable sound prominence that deviates from contextual norm (Jones, 1987). One of the important ways of conveying accent in music is by pitch change, i.e., melodic accent. It is often triggered by change in interval or contour and so is also called interval accent or contour accent (Huron and Royal, 1996). Interval accent most frequently occurs on

the highest pitch after a large interval leap (Graybill, 1989; Lerdahl and Jackendoff, 1983) (Figure 4.1a). The accent can be particularly prominent if the large interval leap is surrounded immediately by stepwise intervals (Graybill, 1989). Contour accent (Figure 4.1b) is proposed to occur at the pivot point where pitch direction changes, especially at the highest pitch of an ascending-descending contour (Thomassen, 1982). Huron and Royal (1996) using a large database with various music styles showed strong support for the pivot accent proposal. Interval accent and contour accent often overlap since the highest pitch after a great interval leap often lies in the pivot position of the melodic contour (Hannon et al., 2004). The degree of melodic accent is proposed to be positively related to the size of pitch interval, i.e., the larger the interval size, the stronger the degree of accent (Lerdahl and Jackendoff, 1983). Nevertheless, so far it is not clear as to how large the interval size should at least be to evoke the perception of melodic accent. Therefore, the above review suggests there does not exist a clearly established threshold for focus perception in Mandarin or melodic accent perception in music. In addition, as introduced in Chapter 1, it is worth testing empirically whether speech and music follow pitch prominence patterns consistent with the prediction of the effort code (Gussenhoven, 2004).



(a) (b)

Figure 4.1 Melodic interval accent (a) and contour accent (b)

Another issue relevant to pitch prominence perception in Mandarin and music is related to the pitch excursion of the post prominence components, especially at the first post-pitch-prominence position. This is because with regard to Mandarin (and many other languages), perception of single focus requires post-focus compression, i.e., the pitch range of post-focus components tend to be compressed in order to highlight the pronounced pitch increase on the focused component (Liu and Xu, 2005; Pierrehumbert and Beckman, 1988; Xu, 1999). In Mandarin, the first post-focus component is usually subject to the greatest extent of distortion because it is right on the downward ramp towards the compressed pitch range (Xu, 1999, 2011) (Figure 4.2). Absence of post-focus compression could lead to the perception of no focus or additional focus (Rump and Collier, 1996). For example, in Dutch a pitch excursion of 2 to 6 semitones on a syllable after the focused syllable could lead to the perception of a double focus (Rump and Collier, 1996). However, it is still unknown in a tonal language such as Mandarin, what the pitch excursion size is for the post-focus components (especially the first component) to be perceived as an additional focus; in other words, it is not known exactly how large the post-focus compression (especially that of the first post-focus component) needs to be for the perception of single focus.

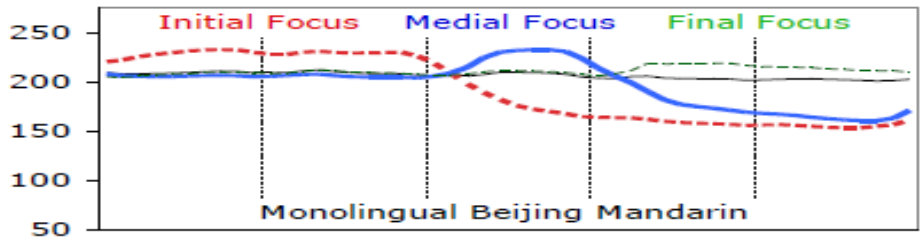


Figure 4.2 Time-normalized mean F0 contours produced by 4 speaker groups. The vertical lines represent syllable boundaries. The solid thin lines represent the no-focus condition (adapted from Xu, 2011).

In terms of music, some have proposed (e.g., Boltz and Jones, 1986) that melodic accent does not necessarily occur on the pivot point as mentioned previously; rather, it can occur on the note immediately following the pivot point, i.e., at the first post-pivot position. This is based on the idea that melodic accent does not have to be associated with the highest pitch of a melodic contour; lower pitch can also convey a sense of accent (Parncutt, 1989). Although this phenomenon is not directly comparable with speech, the common grounds shared by speech and music in this regard is that after the pivot (i.e., the turning point) of a pitch contour of speech or music, there is possibility for pitch prominence to occur at a post-pivot position, either in the form of an additional focus (as in speech) or in the form of a melodic accent (as in music). This further means manipulating the pitch excursion size of the first post-pivot component of the pitch contour of either speech or music can provide clues about the presence or absence of pitch prominence, since in speech post-focused syllables need to be compressed in pitch range otherwise an additional focus will be detected as reviewed above; meanwhile in music the first post-pivot note can be the carrier of melodic accent. This naturally raises the question as to whether speech and music follow the same pitch excursion patterns at the first post-pivot position in signalling pitch prominence. So far there is no empirical research to investigate this issue.

In summary, the above review on speech focus and music melodic accent suggests the following important aspects have not been properly examined: Firstly, there is not a clearly established pitch perception threshold for the perception of focus in Mandarin and music melodic accent. Secondly, with regard to the possibility of post-pivot pitch prominence, it is not known whether speech and music follow the same pitch excursion patterns. Given that both focus in speech and melodic accent in

music signal pitch prominence, and the controversy over the extent to which pitch processing mechanisms are shared between speech and music (Patel, 2008; Peretz, 2012), it is worth further exploring whether or not speech focus and melodic accent follow the same pitch patterns in signalling prominence.

4.1.2 Expectation in speech and music

Expectation is part of psychological laws of mental life responsible for human perception and cognition (Meyer, 1956). More specifically, it is a cognitive mechanism enabling humans to make predictions about the development of future events (Meyer, 1956). Expectation is often reflected in the extent of surprise: A low degree of surprise can reflect consistency with expectation while a high degree can reflect violation of expectation (Reisenzen, 2000; Scherer et al., 2004). In speech prosody, the intonation of surprise is characterized by a large pitch range expansion and a relatively high pitch level (Gussenhoven and Rietvelt, 2000; Lai, 2009). Moreover, prosodically prominent speech elements such as focus and stress are often the main carriers for signaling surprise, as has been evidenced from German (Seppi et al., 2010). Absence of such prosodic cues, e.g., compression or flattening of the pitch contour, could lead to an indication of no surprise or information withdrawal (Gussenhoven, 2004; Lai, 2009).

In music, the degree of surprise is often triggered by different melodic expectation patterns, which have been theorized by Narmour (1990, 1992) in his influential implication-realization (I-R) model of melody. The model is based Meyer's (1956) proposal of musical meaning. Meyer (1956) proposes that musical meaning stems from the way in which listeners' expectations are triggered, impeded or resolved by the musical structures. Particularly, listeners' emotion becomes activated when the

expectation triggered by the preceding musical structure is inhibited or entirely denied by the following musical structure.

Based on Meyer (1956), Narmour (1990, 1992) has proposed a complex Implication-Realisation (I-R) model to account for the perception of musical melody. Following Meyer (1956), Narmour (1990) used "implication" to refer to melodies generating expectations and "realization" as melodies fulfilling expectations. The core idea is that melody perception is built on melodic implications which arise from listeners' expectations for the following melodic events triggered by the preceding events. The events particularly refer to musical intervals.

The principles of the I-R model have been summarized into five key rules for melodic expectation (cf. Krumhansl, 1995a, 1995b): registral direction, intervallic difference, registral return, proximity, and closure. Of particular relevance to this study is the principle of intervallic difference because it is associated with change in pitch range. The principle states that a small preceding interval implies a following interval of similar size, i.e., the same size plus or minus 2 semitones if registral direction changes or the same size plus or minus 3 semitones if the direction stays the same. A large preceding interval implies the following interval of a smaller size, i.e., at least 3 semitones smaller than the large interval if registral direction changes or at least four semitones smaller if registral direction is not changed (Narmour, 1990). This is based on the observation that small intervals tend to be predominant in various music styles (Huron, 2006; Narmour, 1990). A number of studies have used perception and production methods to test the principles of the I-R model. The results on the one hand largely supported the model while on the other hand found the need to include additional factors of tonality (e.g., tonal strength, consonance,

tonal stability, tonal hierarchy) to boost the model's predictive power (Cuddy and Lunney, 1995; Eerola et al., 2002; Krumhansl, 1995a, 1995b; Thompson et al., 1997). The I-R model also has the potential to explain the intonation patterns in speech, as once tentatively outlined in Narmour (1991b). This is because the I-R model is built on the idea that human's expectation patterns are governed by principles that can be applied universally (Narmour, 1990). The principles of the model, therefore, are relevant to all types of melody (e.g., music or speech) (Narmour, 1991b). Indeed, the above review on the pitch patterns of surprise in speech and music suggests that in both domains, small intervals (i.e., small pitch excursions) are generally less likely to trigger surprise than large intervals. The reason could be explained by common motor and perceptual constraints (Patel, 2008). This could serve as further evidence for the close link between speech and music with regard to expectation (Patel, 2008). It is worth pointing out that although pitch in speech does not strictly follow frequency ratios (i.e., semitone intervals) in the same way as music does, research has shown that pitch intervals may indeed be essential to the perception of speech intonation (Hermes, 2006). Evidence can be found in neutral speech (Patel et al., 2006), emotional speech (Curtis and Bharucha, 2010) and stylized interjections (Day-O'Connell, 2013). Moreover, pitch intervals were adopted as a paradigm for examining pitch perception in speech a long time ago (Rietveld and Gussenhoven, 1985; Rump and Collier, 1996). In addition, the use of semitone intervals facilitates cross-modal comparisons between speech and music in terms of pitch processing. Therefore, it is worth testing Narmour's argument (1991b) by empirically examining whether in a tonal (and hence melodic) language like Mandarin, principles of the I-R model such as intervallic difference can be truly applicable in the same way as it is to music.

4.1.3 Research questions

The above review suggests that there could be an intriguing relation between speech and music in terms of pitch prominence and expectation. Nevertheless, some fundamental issues have not been investigated properly, as identified in the above sections. Hence, this study explores the following research questions:

(1) What are the pitch thresholds, if any, for the perception of focus in speech (Mandarin) and melodic accent in music? Do they follow pitch prominence patterns consistent with the prediction of the effort code?

(2) Do speech (Mandarin) and music follow the same pitch excursion patterns in terms of post-pivot pitch prominence?

(3) Is the I-R models' principle of intervallic difference applicable to speech (Mandarin) in the same way as it is to music?

4.2 Experiment 1

Experiment 1 aimed to address research question 1 which is about focus/accent (What is the pitch perception threshold for the perception of focus in Mandarin and melodic accent in music? Do they follow pitch prominence patterns consistent with the prediction of the effort code?) together with question 3 which is about expectation/surprise (Is the I-R models' principle of intervallic difference applicable to Mandarin in the same way as it is to music?). This is because in speech, prosodically prominent elements such as focus are often the main carriers for signalling surprise (Seppi et al., 2010); similarly in music, melodic accents often function to signal musical surprise as well (Jones, 1987). Hence, by making one component in either speech or music prosodically prominent, two research questions

(focus/accent and surprise) can be tackled at the same time. Also note that for research question 3, this study only explores the condition where pitch direction remains unchanged, because surprise in speech usually involves continuous pitch expansion in the same pitch direction rather than the other way round (cf. Kreiman and Sidtis, 2011).

4.2.1 Methods

Participants

15 native Beijing Mandarin speakers with professional musical training background (average training time = 20 years) were recruited as participants (9 females, age *M* = 31 years, *SD* = 3.6). They reported no speech or hearing problems.

Stimuli

Speech

A pre-recorded sentence "*Ta (tone1) xiang (tone3) zuo (tone4) zhe (tone4) dao (tone4) ti (tone2) mu (tone4)*" (He wanted to solve this problem) spoken in a neutral way (i.e., without focus on any syllable) by a native Mandarin Chinese speaker was used as the base sentence. PENTAtrainer1 (Xu and Prom-on, 2010-2015) running under Praat (Boersma and Weenink, 2013) was used to synthetically modify the F_0 contours of the sentence (similar to PSOLA) in such a way that the prosody sounds natural despite the large pitch range modifications. PENTAtrainer1 was based on the PENTA model (Parallel Encoding and Target Approximation) proposed in Xu (2005). The PENTAtrainer1 script was developed from the qTA (quantitative target approximation) implementation (Prom-on et al., 2009) of the PENTA model. The rationale of the model is that pitch contours of tone and intonation can be simulated

as a result of syllable-synchronized target approximation, under the assumption that speech production functions under both biomechanical and linguistic mechanisms (Prom-on et al., 2009). More specifically, the program first extracts for each (manually segmented) syllable an optimal pitch target defined for its height, slope and strength. It then allows the user to arbitrarily modify any of the target parameters and then resynthesize the sentence with the artificial target. Figure 4.3 shows the segmented syllables with the parameters extracted by PENTAtrainer1. For experiment 1, the syllable "*zhe*" (this) was used as the target syllable. Its pitch height parameter (as shown in Figures 4.3 and 4.4) was incrementally raised up to 12 semitones (in one-semitone steps) according to the pitch height of the pre-focused syllable (*zuo*) (more explanation of this is offered below): b= - 8.1384 (the pitch height of *zuo*) + 1 (semitone), + 2 (semitones), + 3 (semitones)…+ 12 (semitones). One semitone was chosen as the step size because a pilot study showed that listeners could not significantly distinguish pitch differences of less than one semitone.

Note that in this study, the pre-focused (*zuo*), focused (*zhe*) and post-focused (*dao*) syllables all have the same falling tone (Tone 4) in Mandarin. Therefore, the pitch manipulation of the focused syllable with reference to the pitch of the pre-focused syllable (as was done in this study) is similar to the pitch manipulation with reference to the pitch of the focused syllable itself and the post-focused syllable respectively. Such design allows the comparison of this study with previous studies on speech focus while at the same time enabling the comparison of speech with music in pitch prominence and expectation: Previous studies on focus perception (in non-tonal languages) manipulated the pitch of focus according to the baseline (i.e. neutral) condition of the focused syllable itself rather than the pre-focused syllable as in this study. While in this study, speech has to be manipulated in the same way as

music (the details are provided in the following section) in order to facilitate comparison between them. This means the component (speech syllable or musical note) should be manipulated according to the pitch of the component immediately preceding the manipulated one (because this is how melodic accent and expectation function in music). Therefore, by making the pre-, on- and post-focused syllables share the same tone (Tone 4), we can guarantee that any of them can serve as the reference (baseline), thus enabling comparisons within this study (speech and music) and across studies (this study and previous studies on speech focus) (cf. Prom-on et al., 2009 for technical details of the extraction of pitch by PENTAtrainer1).

It is also worth mentioning the reason for selecting tone 4 for manipulation is that it produced the clearest pitch target manipulation contour under PENTAtrainer 1 according to our pilot studies. Moreover, the pilot studies showed that listeners' judgement patterns did not differ significantly between stimuli manipulated based on tone 4 and stimuli manipulated based on the rest of the tones (tone 1, 2 and 3).

Music

12 short excerpts in C major were composed for this study (Figure 4.5). Similar to speech, the fourth component (musical note) was the target of manipulation: its pitch height ranged from one semitone above its preceding note all the way to 12 semitones above. Therefore, the target components (syllable or note) in speech and music followed the same manipulation patterns of pitch increase relative to their respective preceding components. This design enables the comparison between speech and music in terms of pitch prominence and expectation.

Note that two different starting tones were used for the melody composition, e.g., ***do*** *re mi fa mi re do* (the first panel of Figure 4.5) and ***re*** *mi fa so fa mi re* (the second

panel of Figure 4.5). The reason is that if we stick to one starting tone (e.g., *do*), then inevitably some of the manipulated notes will be chromatic (i.e., mainly the black keys in the context of C major), for example under the condition where the target note is two semitones above its preceding note (e.g., E-#F). Chromatic tones within C major are highly dissonant and unpleasant (Krumhansl, 1990) and hence would have an impact on listeners' response in terms of melodic expectation. Therefore, in this study two starting tones were used for the stimuli composition to avoid the possible occurrence of chromatic tones.

Each note of the melody was of equal amplitude (56 dB) and was 0.5 second in duration except the last note (which was three times as long as the previous note because it was a dotted half note in time signature 3/4). This was so designed to avoid the possible contribution of intensity and duration to the perception of prominence (accent) (Ellis and Jones, 2009), since the focus of this study was on melodic (pitch) prominence. The total duration of each melody was 4.5 seconds. All melodies were created using Finale 2011 (piano sound).

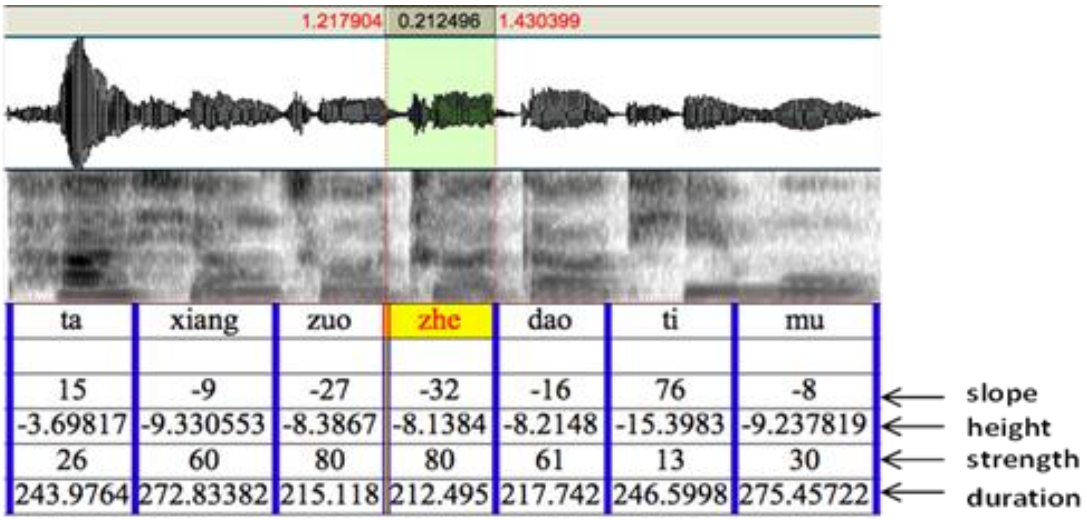


Figure 4.3 The segmentation of the stimulus sentence ("*zhe*" as the target syllable) with parameters automatically derived from PENTAtrainer1 through analysis by synthesis (Xu and Prom-on, 2010-2015).

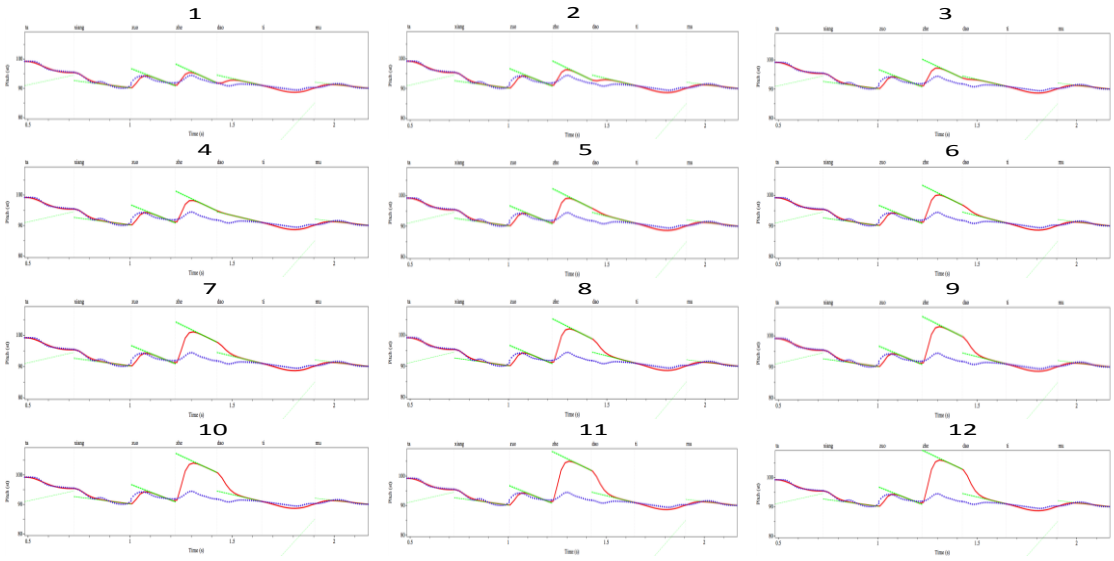


Figure 4.4 The 12 synthesized speech stimuli using PENTAtrainer-1(Xu and Prom-on, 2010-2015). Each stimulus corresponds to a different interval size between the pre-focused syllable *zuo* and the focused-syllable *zhe* (1=1 semitone, 2=2 semitones,… 12=12 semitones). The blue line represents the original speech contour.

The red line represents the synthesized speech contour. The green line represents the pitch target parameters.

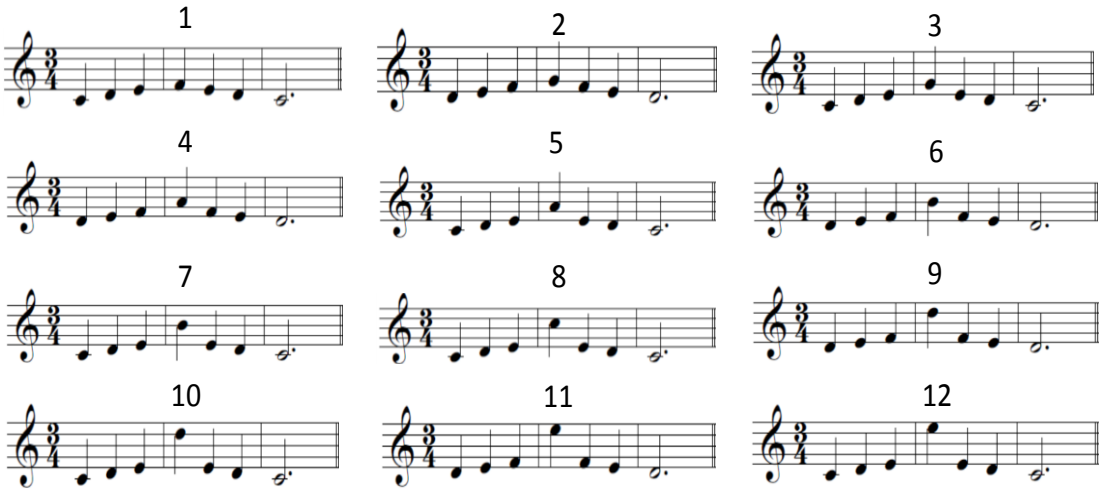


Figure 4.5 The 12 short excerpts composed as the music stimuli. Each excerpt corresponds to a different interval size between the third and fourth note (1=1 semitone, 2=2 semitones,… 12=12 semitones)

Procedure

For the speech experiment, each stimulus sentence was presented three times in a pseudorandom order on a computer. Listeners performed two blocks of tasks: for the first block, they rated the degree of focus conveyed by the syllable "*zhe*" (this) of every sentence on a scale of 1 to 3 (1= no focus; 2 = focus; 3 = a strong degree of focus). They had a fifteen-minute break before starting the second block. The stimuli for the second block were the same as the first block but listeners were asked to rate the degree of surprise conveyed by the syllable "*zhe*" of each sentence on a scale of 1 to 3 (1= not surprising; 2 = surprising; 3 = very surprising). To ensure listeners can distinguish between "focus" and "surprise", different pragmatic contexts were provided. For focus, the context was: He wanted to solve *this* rather than *that*

problem. For surprise, the context was: It was so surprising that he (a very clever student) wanted to solve this problem in an intelligence contest. The problem was so simple that even a not-so-clever student could easily solve, and it turned out that he (with superb intelligence) wanted to solve this problem to show how clever he was.

The music experiment was carried out on a different day. Similar to the speech task, each melody was presented three times in a pseudorandom order on a computer. The same group of listeners participated in the experiment and performed two blocks of tasks: for the first block, they rated the degree of melodic accent conveyed by the fourth note of every melody on a scale of 1 to 3 (1= no melodic accent; 2 = melodic accent; 3 = a strong degree of melodic accent). They had a fifteen-minute break before starting the second block. The stimuli for the second block were the same as the first block but listeners were asked to rate the degree of surprise conveyed by the fourth note of each melody on a scale of 1 to 3 (1= not surprising; 2 = surprising; 3 = very surprising).

4.2.2 Results

Two-way [independent variables: type (speech and music) and interval (12 interval sizes); dependant variable: rating scores] repeated measures ANOVAs showed that for prominence (focus in speech and accent in music), speech and music were not significantly different ($p > 0.05$), but they were significantly different in surprise ($F_{(1, 14)} = 9.2$, $p < 0.01$, $\eta_p^2 = 0.4$). Meanwhile, different interval sizes corresponded to significantly different ratings of prominence ($F_{(11, 154)}=133.4$, $p<0.001$, $\eta_p^2= 0.91$) and surprise ($F_{(11, 154)}=114.8$, $p<0.001$, $\eta_p^2=0.89$). More details are provided below.

Speech

The rusults showed that the larger the interval size, the higher the ratings of the strength of focus (Figure 4.6a) and surprise (Figure 4.6b). This is further confirmed in a one-way repeated measures ANOVA ($F_{(11, 154)}$=168.1, p<0.001, η_p^2 =0.92 for focus; $F_{(11, 154)}$=120.69, p<0.001, η_p^2=0.89 for surprise) where interval size had a significant main effect on the strength of focus and surprise respectively. Furthermore, for focus from 4 semitones onwards (Figure 4.6a) and for surprise from 7 semitones onwards (Figure 4.6b), the average ratings for focus strength and surprise strength respectively were above 2 which is the threshold between no focus/not-surprising (i.e., the rating of 1) and focused/surprising (i.e., the rating of 2). A one-way repeated measures ANOVA further showed that for focus, the difference in ratings between 3 semitones and 4 semitones was significant ($F_{(1, 14)}$ =23.16, p<0.001, η_p^2=0.62) while for surprise, the difference in ratings between 6 semitones and 7 semitones was significant ($F_{(1, 14)}$ =12.51, p=0.003, η_p^2=0.47). This suggests an interval of at least 4 semitones was needed for the perception of focus and that of 7 semitones for the perception of surprise.

Music

For melodic accent, Figure 4.6c shows that the larger the interval size, the higher the degree of accent. This is further confirmed in a one-way repeated measures ANOVA ($F_{(11, 154)}$=30.13, p<0.001, η_p^2 =0.68) where interval size had a significant impact on accent strength. Moreover, Figure 4.6c shows from 3 semitones onwards, the average ratings were above 2 (the threshold between no accent=1 and accent=2) and the difference in ratings between 2 semitones and 3 semitones was significant ($F_{(1,14)}$=24.68, p<0.001, η_p^2 =0.64). This indicates that an interval of at least 3

semitones was needed for the perception of melodic accent. With regard to surprise, the results again showed a significant main effect of interval size on surprise strength ($F_{(11,154)}$=43.09, p<0.001, =0.76). Nevertheless, Figure 4.6d shows that only a partial correlation existed: In the range of 1-7 semitones, the bigger the interval size, the higher the surprise strength and this was especially true from 5 semitones onwards where the average rating was above 2 (the difference between 4 and 5 semitones was significant: $F_{(1,14)}$=8.89, p<0.001, =0.39). However, after 7 semitones, the patterns of surprise strength became more irregular. The surprise strength of 8 semitones was lower than that of 7 semitones and the largest interval (12 semitones) did not correspond to the highest rating of surprise.

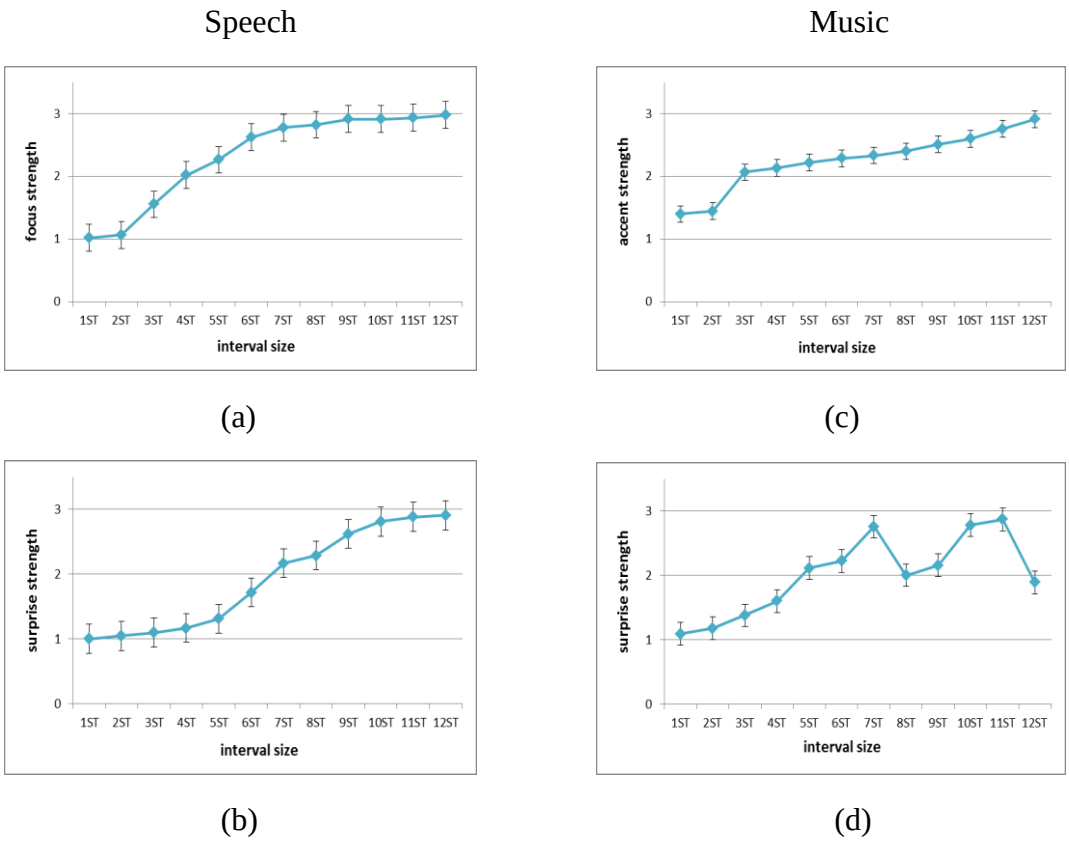


Figure 4.6 The average ratings of focus/accent strength [(a) for speech (c) for music] and surprise strength [(b) for speech (d) for music] for each interval size (ST=semitone).

4.3 Experiment 2

Experiment 2 aimed to address research question 2: Do speech (Mandarin) and music follow the same pitch excursion patterns in terms of post-pivot pitch prominence?

4.3.1 Methods

Participants

The same group of people as experiment 1 participated in experiment 2 (both speech and music tasks).

Stimuli

Speech

The stimulus for manipulation in experiment 2 was one of the synthesized sentences from experiment 1 where the focused syllable "*zhe*" was 12 semitones above its preceding syllable "*zuo*" (as shown in Figures 4.7 and 4.8). In experiment 2, the pitch height of the first post-focus syllable "*dao*" (a classifier modifying its preceding word) was systematically decreased in 12 semitones: b= 3.6113 (the pitch height of *zhe*) -1 (semitone), -2 (semitones), -3 (semitones)…-12 (semitones).

Music

Similar to speech, the music stimulus for manipulation in experiment 2 was a stimulus from experiment 1 where the accented note (the 4th note) was 12 semitones above its preceding note (the 3rd note). In experiment 2, the 5th note (the first post-accented note) was systematically decreased in 12 semitones (one semitone per step) according to the pitch height of the accented 4th note (Figure 4.9). Note that similar to experiment 1, there were two starting notes, e.g., *do* in the second panel of Figure 4.9 and *re* in the first panel of Figure 4.9. The reason is exactly the same as mentioned in experiment 1 (section 4.2.1), i.e., to avoid the occurrence of chromatic

notes which could affect listeners' response and hence become a confound for the experiment.

ta	xiang	zuo	zhe	dao	ti	mu	
15	-9	-27	-32	-16	76	-8	← slope
-3.69817	-9.330553	-8.3867	3.6113	-8.2148	-15.3983	-9.237819	← height
26	60	80	80	61	13	30	← strength
243.9764	272.83382	215.118	212.495	217.742	246.5998	275.45722	← duration

Figure 4.7 The segmentation of the stimulus sentence ("*dao*" as the target syllable) with parameters automatically derived from PENTAtrainer-1 through analysis by synthesis (Xu and Prom-on, 2010-2015).

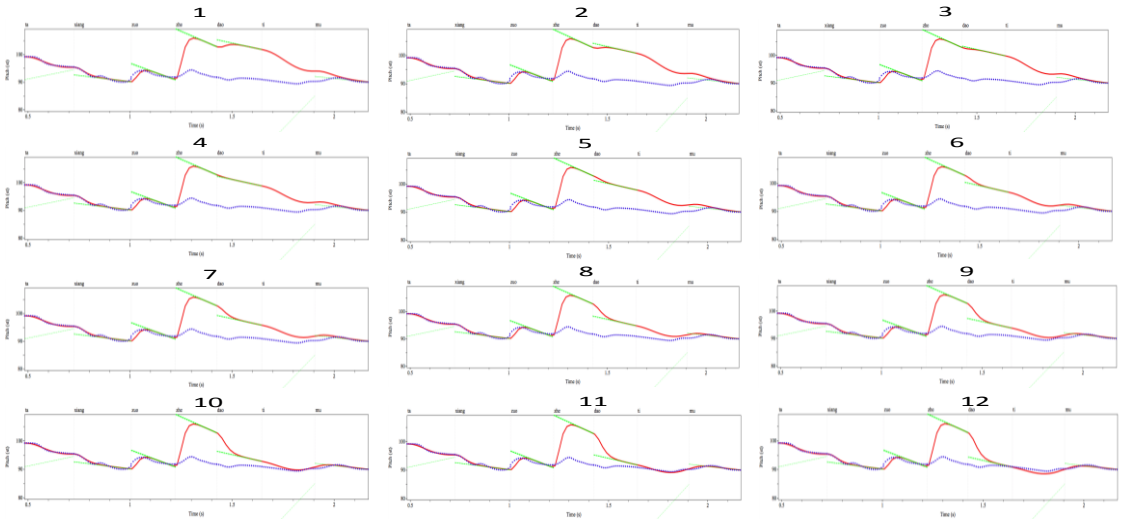


Figure 4.8 The 12 synthesized speech stimuli using PENTAtrainer-1(Xu and Prom-on, 2010-2015). Each stimulus corresponds to a different interval size between the focused syllable *zhe* and the post-focused syllable *dao* (1=1 semitone, 2=2

semitones,… 12=12 semitones). The blue line represents the origin speech contour. The red line represents the synthesized speech contour. The green line represents the pitch target parameters.

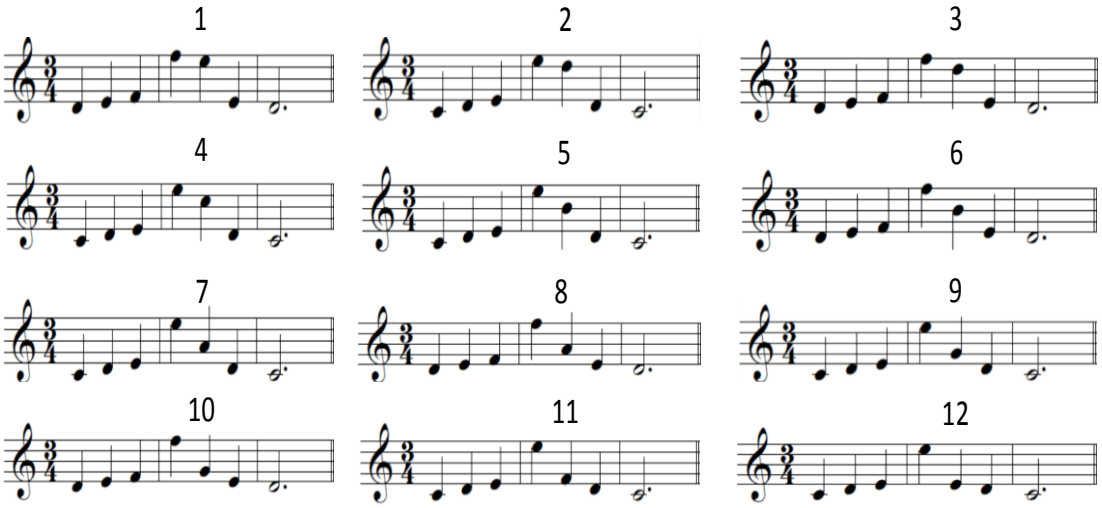


Figure 4.9 The 12 short excerpts of music for experiment 2. Each excerpt corresponds to a different interval size between the fourth and fifth note (1=1 semitone, 2=2 semitones,… 12=12 semitones)

Procedure

Experiment 2 was conducted a week after experiment 1. For the speech experiment, each stimulus sentence was presented three times in a pseudorandom order on a computer. The listeners rated the degree of focus conveyed by the first post-focus syllable "*dao*" of each sentence on a scale of 1 to 3 (1= no focus; 2 = focus; 3 = a strong degree of focus). Similarly, for the music experiment, each melody was presented three times in a pseudorandom order on a computer. The same group of listeners rated the degree of melodic accent conveyed by the 5th note (the first post-accent note) of every melody on a scale of 1 to 3 (1= no melodic accent; 2 = melodic accent; 3 = a strong degree of melodic accent). Note that for the speech experiment,

the design was to mainly ensure that it matched that of the music experiment, especially given the aim of this research discussed in section 4.1.1. Also, such design could indirectly imply whether an additional focus exists or not, which is comparable with previous literature on the perception of additional focus.

4.3.2 Results

A two-way (type: speech and music; interval: 12 interval sizes) repeated measures ANOVA showed that speech and music were significantly different in post-pivot pitch prominence (focus in speech and accent in music) ($F_{(1, 14)}$=27.13, p<0.001, η_p^2 =0.66). Meanwhile, different interval sizes also corresponded to significantly different ratings of prominence ($F_{(11, 154)}$=14.01, p<0.001, η_p^2 =0.5). More details are provided below.

Speech

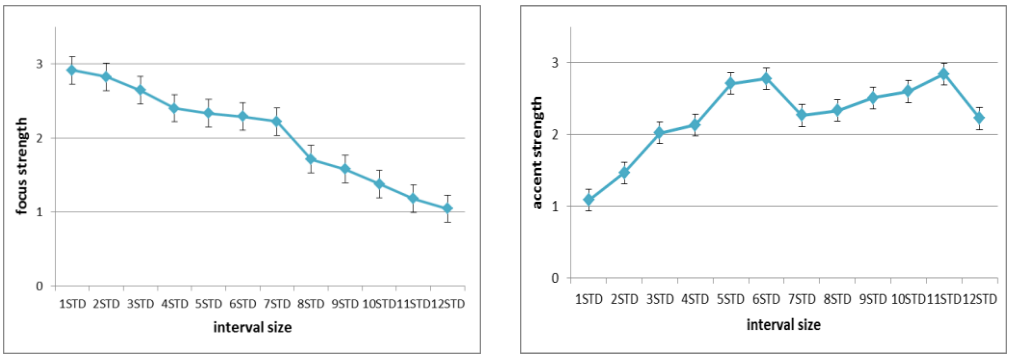
The results showed that the smaller the interval difference, the higher the strength of focus (Figure 4.10a) of the first post-focused syllable. This is further confirmed in a one-way repeated measures ANOVA ($F_{(11, 154)}$=50.31, p<0.001, η_p^2 =0.78), in which interval size had a significant main effect. Moreover, in the range of 1-7 semitones, the first post-focused syllable was perceived as focused, i.e., average ratings were above 2 (the difference between 7 and 8 semitones was significant: $F_{(1, 14)}$=22.77, p<0.001, η_p^2 =0.62).

Music

The results again showed that interval size had a significant main effect on post-pivot accent strength ($F_{(11, 154)}$=24.33, p<0.001, η_p^2=0.64). Furthermore, Figure 4.10b shows that at least an interval of three semitones was needed (above the rating of 2) for the perception of post-pivot accent (the difference between 2 and 3 semitones

was significant: $F_{(1, 14)}$=21.25, p<0.001, η_p^2=0.6). Nevertheless, there does not seem to exist a clear correlation between post-pivot accent strength and interval sizes, especially from 5 semitones onwards: the largest interval (12 semitones) did not correspond to the highest rating of accent strength while relatively small intervals (e.g., 5 or 6 semitones apart) had rather high accent strength, as shown in Figure 4.10b.

Speech Music



(a) (b)

Figure 4.10 The average ratings of focus/accent strength of the first post-pivot component [(a) for speech (b) for music] for each interval size [STD=semitone difference between the pivot component (focus in speech or accent in music) and the first post-pivot component].

4.4 Discussion and conclusion

4.4.1 Pitch prominence in speech and music

In terms of pitch prominence (focus) in speech, the results of experiment 1 showed that generally the strength of focus increased as the pitch excursion size increased, with the threshold lying in 4 semitones, i.e., a pitch excursion of at least 4 semitones was needed to evoke listeners' perception of focus in Mandarin. Non-tonal languages, in contrast, do not require as big an increase in pitch excursion. For instance, in

Dutch an increase of 1.5 or 2 semitones was enough to evoke a perceptual difference in pitch prominence or focus (Rietveld and Gussenhoven, 1985; Rump and Collier, 1996). This could be because tonal languages such as Mandarin use F_0 to signal lexical meanings, and hence more room is needed for F_0 variation to convey lexical meanings in tonal than in non-tonal languages. Correspondingly, the threshold of F_0 to convey other linguistic information such as focus needs to be higher in tonal languages than in non-tonal languages.

With regard to pitch prominence (melodic accent) in music, the results of experiment 1 showed that a pitch increase of 3 semitones was needed to convey melodic accent. As the interval size increased, the perceived strength of melodic accent also increased. The strongest degree of melodic accent appeared at the largest interval leap, i.e., 12 semitones in this study. The results are thus consistent with theoretic proposals that interval size is positively correlated with accent strength, especially in the context of large interval leap (Drake et al., 1991; Lerdahl and Jackendoff, 1983; Monahan et al., 1987).

The results of experiment 1 suggest that speech and music are both similar and different. They are similar because in both domains, high pitch corresponded to a high degree of prominence. This is consistent with previous observation that pitch height is a marker of prosodic prominence in acoustic communications such as speech and music (Patel, 2008; Parncutt, 2003). An acoustic dimension (such as pitch) with high salience usually attracts greater perceptual weight than that with low salience (Benward and White, 1997; ‘t Hart et al., 1990). Therefore, the results of both speech and music experiments are consistent with the prediction of the effort code (Gussenhoven, 2004), i.e., pitch prominence such as focus and accent is always associated with an increase in pitch range. Nevertheless, the results also showed

difference in thresholds for pitch prominence: the threshold of speech focus was one semitone higher than that of music melodic accent. The reason could be that pitch is a fundamental building block in music (Patel, 2008) while in speech less so. This is evidenced from the finding that removing pitch information (i.e., F_0) in speech does not inevitably harm intelligibility, even in a tonal language like Mandarin (Patel et al., 2010). A slight alteration of pitch in music, on the other hand, can easily be heard as "out of tune", a concept that does not apply to speech (Zatorre and Baum, 2012). Therefore, a small change in pitch in music can lead to a significant change in musical meaning (such as melodic accent) while in speech, the magnitude of change in pitch does not need to be as subtle as that in music, even in tonal languages such as Mandarin as shown in this study. Indeed, as argued in Peretz and Hyde (2003), linguistic prosodic contours are often less subtle than music melodic contours, i.e., music has a more fine-grained requirement for pitch compared with speech. Therefore, in music the functional threshold (such as that of pitch prominence) needs to be lower (and hence more subtle) than that in speech.

The results of experiment 2 suggest that speech and music are largely different in terms of post-pivot pitch prominence: in speech, when the lowering of the first post-focused syllable was as small as 1 semitone in relation to the focused syllable, the post-focused syllable can be heard as focused. In music, however, the first post-accented note needed to be at least 3 semitones lower than the accented (pivot) note in order to be perceived as a melodic accent. Moreover, in speech, the focus strength of the post-focused syllable decreased as its pitch gradually lowered whereas in music, the correlation between melodic accent strength and pitch interval size is not as clear as that in speech. The differences between the two domains could be attributed to the fact that firstly, in languages such as Mandarin and English, post-

focus compression is mandatory; a lack of compression or insufficient compression (i.e., the compression range is not big enough) could lead to the perception of an additional focus (Rump and Collier, 1996). Although this study cannot be directly compared with the experiment in Rump and Collier (1996) where the pitch manipulation was not on the first post-focused syllable as done in this study, the present results can be compared with it in an indirect way, i.e., listeners' judgement of the strength of the first post-focus component can imply whether an additional focus exists or not (1=no focus, 2= focused, 3= strongly focused). The results of this study showed that at least a compression size of eight semitones was needed for the perception of single focus, otherwise an additional focus (as indicated by the strength of the first post-focused syllable) was perceived. Therefore, when the compression size was only one semitone, the first post-focused syllable was perceived as an additional focus. In music, however, there is no requirement for pitch range compression of the post-accent components. Correspondingly, the psychological response to pitch patterns in music is not necessarily the same as in speech. As shown in experiment 2, listeners still seemed to follow the three semitone threshold for melodic accent perception as they did in experiment 1 (when the pitch direction went upwards rather than downwards as in experiment 2), i.e., there needed to be a three semitone difference between two notes (the pivot note and first post-pivot note in experiment 2) in order for the second note to be perceived as accented. This is consistent with the view that in music a decrease in pitch can also trigger a sense of accent (Boltz and Jones, 1986).

Another reason for the differences observed above is that unlike speech, music is governed by tonality principles. Melodic accent in music tends to be influenced by tonality factors such as tension (Jones, 1993). A close look at the results of

experiment 2 shows that the strongest melodic accent occurred at the interval of 11 semitones while the largest interval (12 semitones) had a considerably weaker accent. The reason could be attributed to tonal tension which can be triggered by dissonance of intervals, harmonic instability, and melodic attraction (Jones, 1993; Lerdahl and Krumhansl, 2007). An interval of 11 semitones is a major seventh, which is a typical dissonant interval conveying a great sense of bitterness and tension that strongly requires resolution to the tonic (Cooke, 1959). Therefore, such strong degree of tension coupled with a large interval size could qualify as conveying the strongest degree of accent. On the other hand, the 12 semitone interval is a perfect consonance interval (i.e., an octave) with the most stable tonal structure compared with other intervallic combinations (Krumhansl, 1990). Hence, despite its large interval size, it is less likely to trigger tension and consequently accent. Intervals of 6 and 10 semitones are dissonant with a strong degree of tension which requires resolution (Cooke, 1959) and hence convey a relatively strong degree of accent as shown in this study. In addition, the results showed that the perfect fourth (5 semitones) with perfect consonance also had a great degree of accent. This is likely due to the influence of tonal instability (cf. Lerdahl and Krumhansl, 2007): in this study the perfect fourth ended in a tonally unstable note (i.e., *ti*) of the C major, which consequently requires resolution to the tonic, hence giving rise to tension and a strong degree of accent. The reason why such tonality constraint does not occur in experiment 1 could be due to the fact that interval leaps in ascending motion (as in experiment 1) tends to be perceived as more accented than descending interval leaps (as in experiment 2) (Graybill, 1989), and hence interval motion direction could override the influence of tonality on accent perception, as shown in this study. Further research is needed for a deeper understanding of this issue.

In summary, for research question 1, the results of this study showed that in both speech and music, high pitch generally corresponded to a high degree of prominence, which was consistent with the prediction of the effort code. Nevertheless, pitch perception threshold for focus in speech (Mandarin) was one semitone higher than that for the melodic accent in music. For research question 2, speech (Mandarin) and music did not follow the same pitch excursion patterns in terms of post-pivot pitch prominence. The differences between speech and music shown in the two experiments were due to the different functional requirements for pitch in speech and music.

4.4.2 Expectation in speech and music

The results of experiment 1 showed that in both speech and music, small intervals were associated with low degree of expectation violation (i.e., surprise). This is consistent with the I-R models' principle of intervallic difference, especially in terms of music: The principle states that small intervals should be followed by a similarly small sized interval (the same size plus or minus 3 semitones if the pitch direction remains unchanged as in this study) to avoid violation of expectation. The results on music were compatible with the principle because the degree of surprise was very low until the interval of 5 semitones, after which the degree of surprise became significantly large. In this study, 5 semitones were exactly 3 semitones larger than the size of its preceding interval (2 semitones) and hence this part of the results is in line with the intervallic principle. With regard to speech, the results were in the same direction as predicted by the I-R model, i.e., small interval continuation corresponded to low level of surprise. Therefore, the results are consistent with previous findings that in both speech and music, small intervals are preferred (Patel et al., 2006). Such preference for small intervals can be associated with our language

experience (Patel, 2008). This is because greater frequency differences in vocal communication mean larger intervals between pitch targets. According to Fitts's (1954) law, muscular movement is more accurate between short-distance targets (e.g., small pitch intervals) than long-distance targets (e.g., large pitch intervals). Therefore, vocal communication in large frequency difference can be less accurate than that in small frequency difference and is thus less economical in speech articulation. Hence, it is the principle of economy of communication (in speech and music) that leads to the shared preference for small intervals in both domains, and the principle itself could be the results of common motor and perceptual constraints (Patel, 2008).

On the other hand, it is worth noting that although speech was consistent with the direction of the I-R model's prediction, the exact threshold for expectation violation (i.e., surprise) did not fall into the predicted range: in this study, the interval difference between "*xiang*" and "*zuo*" (the interval preceding the manipulated interval) was around 1 semitone, and according to the principle the following interval should be within the range of 1+3=4 semitones in order not to trigger a large extent of surprise. Nevertheless, the results on speech showed that it was from 7 semitones onwards that a large degree of surprise was triggered. Therefore, the results suggest a higher threshold for speech surprise perception than predicted by the I-R model. Moreover, speech had a higher threshold (7 semitones) for violation of expectation than music (5 semitones). The reason for such results is probably that in tonal languages such as Mandarin, pitch serves to differentiate lexical items. Hence, there needs to be enough space for pitch to realize its function as a lexical marker. Consequently, paralinguistic meanings such as surprise have to be allocated to the remaining pitch space. Given the fact that in speech communication pitch

range variation for linguistic information is usually kept small due to the need for economy of articulation (cf. Patel, 2008), the remaining large range of pitch variation is thus allocated to conveying paralinguistic meanings such as surprise. This is also consistent with the findings that surprise intonation usually involves a large pitch excursion and high pitch level (Gussenhoven and Rietvelt, 2000; Lai, 2009). Meanwhile, such inconsistency with the I-R model's prediction also reflects the fact that unlike music, speech does not need to adhere to strict interval ratios to communicate meaning (Zatorre and Baum, 2012) and hence it does not have to strictly follow the intervallic patterns for music as outlined in the I-R model.

In terms of large intervals, speech and music showed significant differences. In speech, large intervals generally corresponded to a large extent of surprise (which is consistent with the I-R model). With regard to the results on music, there is not a direct correlation between interval size and the degree of surprise in the range of large intervals (from 8 semitones onwards): For example, the interval of 8 semitones had a weaker degree of surprise than 7 semitones; the interval of 12 semitones was weaker in surprise than the intervals of 10 and 11 semitones. This pattern is inconsistent with the prediction of the I-R model, and indeed as previous studies (cf., Krumhansl, 1995b) have noted, the principle of intervallic difference sometimes has the weakest predictive power due to the influence of additional factors such as tonal stability. More specifically, previous studies (Eerola, 2002; Krumhansl, 1995b; Thompson et al., 1997) have reported that tonally less stable notes are generally perceived as more surprising than tonally stable notes. In this study, the 7 semitone interval ended in *ti* (the leading note) which is the least stable note in C major due to its inclination to resolve to the tonic *do*. This could lead to a high degree of surprise. In contrast, the 8 semitone interval ended in *do*, which is the tonic of the musical key

it is situated in (C major). It is the most stable note (Meyer, 1956) and is therefore less surprising than the leading note. The 12 semitone interval, despite being the largest interval, was rated less surprising than smaller intervals (e.g., 10 and 11 semitones). The reason is that it ended in *mi* which is the median of C major (the musical key it is situated in). Since the median is the third most stable note of a musical key (after the tonic and the dominant, cf. Meyer, 1956), it is consequently less surprising, especially when compared with intervals of 10 and 11 semitones (the minor and major seventh) which require resolution to the tonic and hence less stable (Meyer, 1956). Since such tonal stability exists only in music rather than in speech, it is not surprising that the results showed different pitch expectation patterns of speech from those of music.

In summary, for research question 3, the results suggest that in terms of small intervals, speech (Mandarin) and music were similar in the sense that both were consistent with the prediction of the I-R model: small intervals were preferred over large intervals to avoid expectation violation (e.g., surprise). Nevertheless, the model could not predict the exact pitch threshold for surprise in speech (which was higher than music). In addition, in terms of large intervals, music was noticeably different from speech due to constraints from factors such as tonal stability which has no counterpart in speech.

In conclusion, this study empirically examined previously unexplored yet fundamental aspects of pitch processing in speech and music: pitch prominence (i.e., focus in speech and melodic accent in music) and melodic expectations (i.e., the degree of surprise) within the framework of the I-R model. The results suggest that there can be some extent of overlap between speech and music in terms of pitch prominence (e.g., high pitch corresponded to great prominence) and expectation

patterns (e.g., small intervals were preferred over large intervals). Nevertheless, the differences seemed to have outweighed the similarities between the two domains due to functional differences of pitch in speech and music. Therefore, in terms of the two views regarding the relations between speech and music melody introduced in section 4.1, the results are more in favour of the second view: speech and music melody tend to require specialized pitch patterns unique to their own respective communication purposes (Peretz, 2006, 2012; Zatorre and Baum, 2012), although whether they are governed by separate neural mechanisms is still a topic of ongoing debate. Of course, this does not negate the commonalities between them (Peretz, 2012), as the results of this study also showed similar tendencies shared between the two domains. This naturally raises the question about exactly to what extent speech and music melody can be related to each other. More research along this line would provide a more comprehensive answer to this question and hence sheds new light on comparative studies on pitch processing of speech and music.

Chapter 5

Conclusion

5.1 Summaries of the main findings of this thesis

This thesis examined the relations between music and speech from the perspectives of dynamics, timbre and pitch using production (for dynamics) and perception (for timbre and pitch) methods. Chapter 2 dealt with dynamics: unlike previous research, in this thesis finger force and articulatory effort were used as indexes reflecting the dynamics of affective piano performance and speech production, respectively. Moreover, for the first time physical constraints such as piano fingerings and speech articulatory constraints were included due to their potential contribution to different patterns of dynamics. A piano performance experiment and speech production experiment were conducted in four emotions: anger, fear, happiness and sadness. The results showed that in both piano performance and speech production, anger and happiness generally had high dynamics while sadness had the lowest dynamics, which can be interpreted from an evolutionary perspective. Fingerings interacted with fear in the piano experiment and articulatory constraints interacted with anger in the speech experiment, i.e., large physical constraints produced significantly higher dynamics than small physical constraints in piano performance under the condition of fear and in speech production under the condition of anger. In addition, the results also showed that affective speech production on the whole had higher dynamics than affective piano performance, which may be due to the biomechanical differences between speech articulators and fingers. Using production experiments, this is the first study to show quantitative evidence for the importance of considering motor

aspects such as dynamics in comparing similarities and differences between music and speech.

Chapter 3 dealt with timbre, with a focus on the emotional connotations of musical timbre of isolated instrument sounds through the perspective of affective speech using behavioural and ERP experiments. The behavioural experiment compared the timbre (i.e., voice quality) of affective speech and the timbre of isolated instrument sounds categorized by listeners into three emotions: anger, happiness and sadness. The results showed that there were no significant differences between affective speech and musical instruments in terms of the timbral acoustic features in each category of the emotions, suggesting that the timbre of musical instrument sounds in each emotion (anger, happiness and sadness) is acoustically similar to the timbre of affective speech of the same emotional category. Two ERP experiments were conducted to further explore the neural processing of affective speech and instrument sounds. The first one tested the ERP patterns (the P200 and LPC) of affective speech and instrument sounds separately. The results showed that overall, speech had significantly higher P200 and LPC amplitude than isolated instrument sounds, which was probably due to the brain processing advantage of human voice. Nevertheless, similarities also existed: in both speech and instrument conditions, anger was higher than happiness and sadness in the P200 and LPC amplitude; sadness was higher than happiness in the LPC amplitude. The second ERP experiment used a priming paradigm, with isolated instrument sounds as primes and affective speech as targets. The results showed that emotionally incongruent instrument-speech pairs triggered larger N400 than emotionally congruent pairs. Taken together, this is the first empirical study to show that even simple, isolated musical instrument sounds can convey emotional connotations in a way similar to affective speech. In addition, the

timbral features of both instruments and speech in each emotional category were consistent with the prediction of body-size projection theory on emotion.

Chapter 4 empirically examined previously unexplored yet fundamental aspects of pitch processing in speech and music: pitch prominence (i.e., focus in speech and melodic accent in music) and melodic expectations (i.e., the degree of surprise) within the framework of the implication realization (I-R) model. The results showed some degree of overlap between speech and music: high pitch generally corresponded to a high degree of prominence; small intervals were preferred over large intervals to avoid expectation violation (i.e., surprise), which was consistent with the prediction of the I-R model. Nevertheless, the differences seemed to have outweighed the similarities due to functional differences of pitch in speech and music: the pitch perception thresholds for pitch prominence and surprise were higher in speech than in music; speech and music did not follow the same pitch excursion patterns in terms of post-pivot pitch prominence; with regard to pitch expectation patterns in the range of large intervals, music was noticeably different from speech due to constraints from factors such as tonal stability which has no counterpart in speech. Therefore, the results have provided new evidence for the view (Peretz, 2006, 2012; Zatorre and Baum, 2012) that speech and music melody tend to require specialized pitch patterns unique to their own respective communication purposes, although commonalities also exist.

5.2 Relations between music and speech

5.2.1 Music and speech: overlapping

On the one hand, the results of this thesis suggest that music and speech are overlapping in some aspects, particularly in terms of dynamics and timbre: dynamics

in both music and speech were high in anger and happiness while low in sadness; timbre of isolated musical instrument sounds conveyed emotional connotations in a way similar to affective speech; high pitch corresponded to great prominence.

This thesis has shown that the overlapping aspects between music and speech as mentioned above can be interpreted from a bio-evolutionary perspective. In terms of dynamics and timbre, the results are consistent with evolutionary accounts on emotion: anger corresponded to high dynamics in speech and music because evolutionarily anger, like all other emotions, originates from natural selection pressure on animals (Darwin, 1872). The evolutionary function of anger is to facilitate the fight or attack response in situations that threaten survival (Darwin, 1872). High dynamics in anger (as in angry speech and piano performance shown in this thesis) reflects great physical strength and energy triggered by the fight response (Xu et al., 2013a). Furthermore, anger was also found in this thesis associated with rough timbre in music and speech. Rough sound timbre, meanwhile, correlates to large body size (e.g., lions) due to simple physical laws (Morton, 1977). In a physical confrontation (i.e., fight), a large body size stands at a better chance of beating off enemies. Therefore, anger carries the evolutionary message of a large body size projection (Xu et al., 2013a) and hence the rough sound timbre of angry musical sounds and speech is a reflection of this message. Happiness, on the other hand, is to signal appeasement or associability (Morton, 1977) as well as willingness to play (Panksepp, 2005). Hence, it is beneficial for sound signalers to produce highly vigorous (i.e., dynamic) sounds in conveying happiness so that they can be audible to potential mates and listeners (Xu et al., 2013a), as reflected in happy speech and piano performance in this thesis. In addition, in both music and speech, happiness was associated with tone-like sound quality, which correlates with small

body size (e.g., birds) according to physical laws (Morton, 1977). Evolutionarily, small body size signals lack of threat and social attractiveness (Morton, 1977; Xu et al., 2013b). Therefore, the tone-like timbre of happy musical sounds and speech both project small body size and hence send an inviting signal. With regard to sadness, the low dynamics in both piano performance and speech production reflected low physical activation level, which could evolutionarily imply a tendency to beg for sympathy (Shaver et al., 1987). The differences in timbre between sad instrument sounds and sad speech were due to two different kinds of sadness: grieving sadness and depressed sadness (Xu et al., 2013a), which have different body size projections. Grieving sadness projects a large body size due to its demanding nature, while depressed sadness has relatively neutral size projection due to its lack of communicative intention (Xu et al., 2013a). This was reflected in the different spectrum patterns of sad speech (depressed sadness) and sad instrument sounds (grieving sadness) in this thesis. In terms of fear, the relatively high dynamics in piano performance and speech production is consistent with the argument that fear can be an evolutionarily defensive strategy (LeDoux, 1996). This is evidenced from animal alarm calls (when in fear) which is usually used as a useful antipredator defensive strategy for the sake of group survival (Caro, 2005). To serve this purpose, alarm calls (i.e.., fearful vocalizations) should be reasonably high in dynamics (i.e., vigorousness). Similarly, production of musical excerpts of fear could also be highly dynamic, analogous to human fearful speech or animal alarm calls. With regard to pitch prominence, this thesis showed that in both music and speech, high pitch corresponded to a high degree of prominence. This is consistent with the prediction of the effort code (Gussenhoven, 2004), which can explain pitch prominence in both speech and music from a biological perspective (Cross and Woodruff, 2009).

Such cross-domain similarities in dynamics, timbre and pitch between music and speech from a bio-evolutionary perspective as shown in this thesis echo previous proposals (e.g., Darwin, 1871; Cross, 2009a) that both music and speech can have bio-evolutionary implications. It is worth mentioning that while there is sufficient support for the evolutionary implications of speech, which has been evidenced from anatomy of the human vocal tract, vocal learning, and neurobiology of language acquisition (cf. Patel, 2008), it is still a matter of debate whether music has any evolutionary bearings at all, i.e., whether music has been specifically adapted as an independent trait as put forth in Darwin (1871). Admittedly, music is unlike eating, drinking and speech which are indispensable for normal functioning in human society (Pinker, 1997). Nevertheless, the evolutionary implications of music may be clearer if viewed from the perspective of evolutionary fitness, i.e., the impact of music on humans' personal development and social bonding (Cross, 2009b). For example, similar to speech, music is an important way of enhancing neural adaptation and brain elasticity (Huttenlocher, 2002), i.e., it provides a platform for practicing intellectual abilities (e.g., note reading) and physical exercise (e.g., singing or dancing to the tune), during the process of which the cognitive abilities of individuals are exercised and fostered (Cross, 2003). Neural imaging studies have also provided sufficient support for the role music plays in changing brain structures by enlarging certain brain areas as a result of musical training and experience (e.g., Münte et al., 2002; Pantev et al., 1998). In terms of social bonding, speech is not the only medium through which people interact with each other; music also plays an important role in strengthening social bonds through situations such as ritual ceremonies and collective music making (Cross, 2009a, 2009b; Kogan, 1997). This is because musical activity is by nature cultural and hence inherently has the

potential to unify members of society (Kogan, 1997). The mood-regulating function of music also prepares for the transmission of individual mood to collective mood during the music making process (Sloboda and O'Neill, 2001), and hence music could evolutionarily facilitate the regulation of individuals and members of society as a whole. Therefore, many theories are pointing to the direction that music, like speech, has evolutionary implications. This thesis has further shown that through the platform of emotion, music and speech can convey similar evolutionary signals, thus providing further evidence for the aforementioned line of argument.

5.2.2 Music and speech: distinct

On the other hand, the results of this thesis also imply considerable differences between music and speech: speech production overall had higher dynamics than piano performance; affective speech and music performance interacted differently with physical constraints; there was a significant brain processing advantage for speech over musical instrument sounds as evidenced from the ERP results; the use of pitch (in terms of prominence and expectation) was largely different in speech than music. The reasons could be attributed to the fact that although both music and speech can have evolutionary implications as discussed in the above section, music does not have an equal status with speech in evolution, particularly with regard to communication which is a crucial aspect of evolutionary adaptation. More specifically, music is not as efficient as speech for communication, i.e., speech can be the evolutionary product primarily selected for communication while music less so. This is evident from the fact that language and speech exist fundamentally for communication purposes while music, although also communicates meaning, is more of an art form that conveys aesthetic messages (Patel, 2008). From a bio-physiological perspective, this partly explains why oral-facial muscles have evolved

to have fast-twitch fibers and motor protein such as myosin that enable fast acceleration and rapid speech to meet different levels of speech demand (Burke, 1981; Williams and Warwick, 1980). Finger muscles (as in the context of piano performance reported in Chapter 2), on the other hand, do not contract as fast as speech articulatory muscles (Gentil and Tournier, 1998) due to the presence of long tendons, joints and muscle mass between fibers and skeletal joints. Consequently it is not surprising that speech production had higher dynamics than music performance as shown in this thesis. With regard to timbre, speech triggered larger ERP amplitude than instrument sounds. This is likely because compared with music, speech is more relevant to human communication and hence has developed a processing advantage in the human brain (Latinus and Belin, 2011). Deficits in speech perception can seriously hamper personal development because humans by nature need to be sensitive to speech sounds in order to adapt and survive in society (Kuhl, 1988) while in contrast, deficits in music perception (e.g., tone-deafness) are much less of a threat to personal development. This can be further evidenced from the fact that people suffering from music tone-deafness can still perceive linguistic intonation relatively well (Peretz and Hyde, 2003). Hence, the overall higher dynamics in speech production than in piano performance and the brain's processing advantage for human speech shown in this thesis reflects the more direct role of speech in human communication than music. In addition, the different use of pitch in music than in speech as shown in this thesis provides further insight into the functional differences between music and speech: music is more relevant to serving aesthetic purposes than speech. The constraints of music tonality in both melodic accent and expectation reflect the aesthetic strategies in melodic composition (Narmour, 1991a). Speech, in contrast, does not need as much tonality constraint as

does music for aesthetic purposes due to the primary role of speech for communication. A vivid example (cf. Patel, 2008) is that we seldom hum to linguistic intonation contours; instead, we usually hum to music tunes because of the aesthetic pleasure the tunes bring to us. Therefore, the differences between music and speech shown in this thesis suggest that although both of them have evolutionary implications and can be overlapping in many aspects, music and speech still maintain distinct functions unique to their respective domains.

5.3 Implications for exploring music-speech relations and future directions

In a nutshell, from three perspectives (dynamics, timbre and pitch) essential to music and speech, this thesis has shed new light on the overlapping yet distinct relations between the two domains. The implications of such in-depth exploration of the relations between music and speech are that firstly, a well-established link between the two could facilitate theory formulation in which the principles that apply to speech could be used to explain music or vice versa (Juslin and Laukka, 2003). Moreover, the formation of such a link between the two domains could shed further light on the origins of music and language (Darwin, 1871; Brown, 2000). Thirdly, it could help enhance our understanding of the functional and neural characteristics of music and speech (Patel and Peretz, 1997), thus contributing to a more comprehensive understanding of both domains. Fourthly, it also facilitates the exploration into neural plasticity since both music and speech are important ways of mediating the function and structure of the brain (Slevc, 2012; Asaridou and McQueen, 2013). Such cross-domain investigation could be especially useful for studies on children's mental development due to the adaptive functions of both speech and music (Fritz et al., 2013).

Future research could examine the transfer effect from music to speech and vice versa. While the effect of music on speech processing has been relatively intensively investigated, the evidence for language facilitating music cognition has been mixed so far and is not as compelling as that for music to language transfer (Slevc, 2012; Asaridou and McQueen, 2013). Another line of research would be to explore the link between bimusicality (i.e., early exposure to more than one musical culture) and bilingualism (Slevc, 2012). In addition, using neurophysiological approaches (fMRI, EEG, MEG, etc.) to further examine the parallels and dissociations between the sensory-motor aspects of music and speech could shed more light on human sensory-motor skills in general, with important clinical implications as well (Zatorre, 2013). All in all, systematic and in-depth explorations of the relations between music and speech can broaden our understanding of human nature in a wider sense and deepen our insight into human's unique ability to construct and appreciate the rich and complex sound landscape of the human world.

References:

Adams, S. G., Weismer G., and Kent, R. D. (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech and Hearing Research*, 36, 41-54.

Agus, T. R., Suied, C., Thorpe, S. J., and Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *Journal of the Acoustical Society of America*, 131, 4124–4133.

Alluri, V., and Toiviainen, P. (2010). Exploring perceptual and acoustic correlates of polyphonic timbre. *Music Perception*, 27 (3), 223–241.

Asaridou, S. S., and McQueen, J. M. (2013). Speech and music shape the listening brain: Evidence for shared domain-general mechanisms. *Frontiers in Psycbology*, 4, 321.

Askenfelt, A. (1991). Voices and strings: Cousins or not? In J. Sundberg, L. Nord, and R. Carlson (Eds.), *Music, language, speech and brain: Proceedings of an international symposium at the Wenner-Gren Center, Stockholm* (Vol. 59, pp. 243–259). London: Macmillan Press.

Austin, G. J., Leslie, B. M. and Ruby, L. K. (1989). Variations of the flexor digitorum superficialis of the small finger. *Journal of Hand Surgery*, 14A, 262-267.

Baker, D. S., Gaul, J. S. Jr, Williams, V. K. and Graves, M. (1981) The little finger superficialis-clinical investigation of its anatomic and functional shortcomings. *Journal of Hand Surgery*, 6, 374-378.

Balkwill, L.-L., and Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: psychophysical and cultural cues. *Music Perception*, 17, 43–64.

Balkwill, L.-L., Thompson, W. F., and Matsunaga, R. (2004). Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners. *Japanese Psychological Research*, 46(4), 337–349.

Bamberger, J. (1976). The musical significance of Beethoven's fingerings in the piano sonatas. *Music Forum*, 4, 237-280.

Banse, R., and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70, 614–636.

Benward, B., and White, G. (1997). *Music in theory and practice* (Vol.1, 6th ed). Madison, WI: Brown and Benchmark.

Bernays, M., and Traube, M. (2012). Piano touch analysis: A MATLAB toolbox for extracting performance descriptors from high resolution keyboard and pedalling data.

In *Proceedings of the International Symposium on Performance Science (ISPS2011)* (pp. 299–304). Utrecht, Netherlands: European Association of Conservatoires.

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., and Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion*, 19, 1113–1139.

Boersma, P., and Weenink, D. (2013). *Praat: Doing phonetics by computer* [Computer program]. Version 5.3.59, retrieved from http://www.praat.org/

Boiten, F. A., Frijda, N. H., and Wientjes, C. J. E. (1994). Emotions and respiratory patterns: Review and critical analysis. *International Journal of Psychophysiology*, 17, 103-28.

Bolinger, D. (1985). *Intonation and its parts: Melody in spoken English*. London: Edward Arnold.

Boltz, M. G., and Jones, M. R. (1986). Does rule recursion make melodies easier to reproduce? If not, what does? *Cognitive Psychology*, 18, 389 – 431.

Boulez, P. (1987). Timbre and composition – timbre and language. *Contemporary Music Review*, 2, 161–71.

Bresin, R., and Friberg, A. (2000). Emotional coloring of computer-controlled music performances. *Computer Music Journal*, 24, 44–63.

Brown, S. (2000). ‘‘The musilanguage’’ model of music evolution. In N. Wallin, B. Merker, and S. Brown (Eds.), *The origins of music* (pp. 271–300). Cambridge: A Bradford Book: The MIT press.

Bruce, G. (1977). *Swedish word accents in sentence perspective.* Lund: Lund University Press.

Buck, R. (1984). *The communication of emotion*. New York: Guilford Press.

Burke, R. E. (1981). Motor units: anatomy, physiology and functional organization. In Brooks (ed.), *Handbook of Physiology. Section 1: The Nervous System. Vol. II. Motor Control, Part I* (pp.345-422). Washington, DC: American Physiological Society.

Carlo, N. S., and Guaitella, I. (2004). Facial expressions of emotion in speech and singing. *Semiotica*, 149 (1/4), 37–55.

Caro, T. (2005). *Antipredator defenses in birds and mammals.* Chicago, Illinois: University of Chicago Press.

Charest, I., Pernet, C. R., Rousselet, G.A., Quinones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J. P., and Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, 10, 127.

Chen, Y., and Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, 36, 724–746.

Cheng, C., and Xu, Y. (2013). Articulatory limit and extreme segmental reduction in Taiwan Mandarin. *Journal of the Acoustical Society of America*, 134(6), 4481–4495.

Cheng, C., and Xu, Y. (in press). Mechanism of disyllabic tonal reduction in Taiwan Mandarin. *Language and Speech.*

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. and Maneewongvatana, S. (2008). Encoding emotions in speech with the size code—A perceptual investigation. *Phonetica*, 65, 210-230.

Clark, J., and Yallop, C. (1990). *An introduction to phonetics and phonology*. Oxford, UK: Blackwell.

Clarke, E . F., Parncutt, R., Raekallio, M., and Sloboda, J. A. (1997). Talking fingerings: An interview study of pianists' views on fingering. *Musicae Scientiae*, 1, 87-107.

Cooke, D. (1959). *The language of music*. Oxford: Oxford University Press.

Cooper, W., Eady, S., and Mueller, P. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, 77(6), 2142–2155.

Cross, I. (2003). Music, cognition, culture and evolution. *Contemporary Music Review*, 22, 79–89.

Cross, I. (2009a). The nature of music and its evolution. In S. Hallam, I. Cross and M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 3-13). Oxford: Oxford University Press.

Cross, I. (2009b). The evolutionary nature of musical meaning. *Musicae Scientiae*, 13, 179–200.

Cross, I., and Woodruff, G. E. (2009). Music as a communicative medium. In R. Botha & C. Knight (Eds.), *The prehistory of language* (Vol. 1, pp. 113–144). Oxford, England: Oxford University Press.

Cross, I., Fitch, W. T., Aboitiz, F., Iriki, A., Jarvis, E. D., Lewis, J., Liebal, K., Merker, B., Stout, D., and Trehub, S. E. (2013). Culture and evolution. In M. A. Arbib (Ed.), *Language, music, and the brain: A mysterious relationship* (pp. 541-562). Cambridge, MA: MIT Press.

Cuddy, L. L., and Lunney, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgements of continuity. *Perception and Psychophysics*, 57(4), 451–462.

Cummings, A., Ceponiene, R., Koyama, A., Saygin, A., Townsend, J., and Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain Research*, 1115, 92–107.

Curtis, M. E., and Bharucha, J. J. (2010). The minor third communicates sadness in speech, mirroring its use in music. *Emotion*, 10, 335–348.

Darwin, C. (1871). *The descent of man, selection in relation to sex*. London: Pickering and Chatto.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London, England: John Murray.

Day-O'Connell, J. (2013), Speech, song, and the minor third: an acoustic study of the stylized interjection, *Music Perception*, 30, 441–462.

de Jong, K. (2004). Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *Journal of Phonetics*, 32, 493–516.

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, 134, 9–21.

Drake, C., Dowling, W. J., and Palmer, C. (1991). Accent structures in the reproduction of simple tunes by children and adult pianists. *Music Perception,* 8, 315–334.

Dunn, B. R., Dunn, D. A., Languis, M., and Andrews, D. (1998). The relation of ERP components to complex memory processing, *Brain and Cognition*, 36, 355–376.

Edwards, J. R., Beckman, M. E., and Fletcher, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America, 89*, 369-382.

Eerola, T., Ferrer, R., and Alluri, V. (2012). Timbre and Affect Dimensions: Evidence from Affect and Similarity Ratings and Acoustic Correlates of Isolated Instrument Sounds. *Music Perception*, 30(1), 49−70.

Eerola, T., Toiviainen, P., and Krumhansl, C. L. (2002). Real-time prediction of melodies: Continuous predictability judgements and dynamic models. In *Proceedings of the Seventh International Conference on Music Perception and Cognition* (pp.473–476). Adelaide, Australia: Causal Productions.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6, 169-200.

Ellis, R. J., and Jones, M. R. (2009). The role of accent salience and joint accent structure in meter perception. *Journal of Experimental Psychology. Human Perception and Performance*, 35(1), 264-280.

Erickson, D., Iwata, R., Endo, L., and Fujino, A. (2004). Effect of tone height on jaw and tone articulation in Mandarin Chinese. In *Proceedings of the International Symposium on Tonal Aspects of Languages* (pp.53-56). Beijing, China.

Feng, C., Wang, L., Liu, C., Zhu, X., Dai, R., Mai, X., and Luo, Y. J. (2012). The time course of the influence of valence and arousal on the implicit processing of affective pictures. *PLoS One*, 7, e29668.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47 (6), 381–391.

Fonagy, I., and Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift für Phonetik*, 16, 293–326.

Frey, A., Aramaki, M., and Besson, M. (2014). Conceptual priming for realistic auditory scenes and for auditory words. *Brain and Cognition*, 84 (1), 141–152.

Fritz, J., Poeppel, D., Trainor, L., Schlaug, G., Patel, A., Peretz, I., Rauschecker, J., Halle, J., Stregapede, F., and Parsons, L. (2013). The neurobiology of language, speech and music. In M. A. Arbib (Ed.), *Language, music, and the brain: A mysterious relationship* (pp. 417-459). Cambridge, MA: MIT Press.

Gabrielsson, A. (1995). Expressive intention and performance. In R. Steinberg (Ed.), *Music and the mind machine* (pp. 35-47). New York: Springer.

Gabrielsson, A. (2001). Emotions in strong experiences with music. In P. N. Juslin, and J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 431–449). Oxford: Oxford University Press.

Gabrielsson, A. (2003). Music performance research at the millenium. *Psychology of Music*, 31, 221–272.

Gentil, M., and Tournier, C. L. (1998). Differences in fine control of forces generated by the tongue, lips and fingers in humans. *Archives of Oral Biology*, 43, 517–523.

Giordano, B. L., and McAdams, S. (2010). Sound source mechanics and musical timbre perception: Evidence from previous studies. *Music Perception*, 28(2), 155–168.

Gobl, C., and Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40, 189–212.

Goerlich, K. S., Witteman, J., Schiller, N. O., Van Heuven, V. J., Aleman, A., and Martens, S. (2012). The nature of affective priming in music and speech. *Journal of Cognitive Neuroscience*, 24, 1725 –1741.

Goudbeek M. and Scherer K. R. (2010). Beyond arousal: valence and potency/control cues in the vocal expression of emotion. *Journal of the Acoustical Society of America*, 128, 1322–1336.

Goydke, K. N., Altenmüller, E., Möller, J., and Münte, T. F. (2004). Changes in emotional tone and instrumental timbre are reflected by the mismatch negativity. *Brain Research*, 21, 351–359.

Graybill, R. (1989). Phenomenal accent and meter in the species exercise. *In Theory Only*, 11, 11–43.

Griffiths, T. D., and Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5, 887–92.

Grigor, J., Van Toller, S., Behan, J., and Richardson, A. (1999). The effect of odour priming on long latency visual evoked potentials of matching and mismatching objects. *Chemical Senses*, 24, 137–144.

Grillner, S., Lindblom, B., Lubker, J., and Persson, A. (1982). *Speech Motor Control*. New York, NY: Raven Press.

Grossmann, T., Oberecker, R., Koch, S. P., and Friederici, A. D. (2010). The developmental origins of voice processing in the human brain. *Neuron*, 65, 852–858.

Gunter, G. S. (1960). Traumatic avulsion of the insertion of flexor digitorum profundus. *The Australian and New Zealand Journal of Surgery*, 30, 1-9.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

Gussenhoven, C., and Rietveld, T. (2000). The behavior of H and L under variations in pitch range in Dutch rising contours. *Language and Speech*, 43(2), 183–203.

Hamm, J. P., Johnson, B.W., and Kirk, I. J. (2002). Comparison of the N300 and N400 ERPs to picture stimuli in congruent and incongruent contexts. *Clinical Neurophysiology*, 113, 1339–1350.

Hannon, E. E., Snyder, J. S., Eerola, T., and Krumhansl, C. L. (2004). The role of melodic and temporal cues in perceiving musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, 30, 956–974.

Hanssen, J., Peters, J., Gussenhoven, C. (2008). Prosodic effects of focus in Dutch declaratives. In *Proceedings of the 4th International Conferences on Speech Prosody* (pp.609-612). Campinas.

Hermes, D. J. (2006). Stylization of pitch contours. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schließer (Eds.), *Methods in empirical prosody research* (pp. 29–61). Berlin Germany: de Gruyter.

Hertrich, I., and Ackermann, H. (1997). Articulatory control of phonological vowel length contrasts: kinematic analysis of labial gestures. *Journal of the Acoustical Society of America*, 102, 523–536.

Hinojosa, J. A., Carretié, L., Méndez- Bértolo, C., Míguez, A., and Pozo, M.A. (2009). Arousal contributions to affective priming: electrophysiological correlates. *Emotion,* 9, 164–171.

Holmes, P. A. (2011). An exploration on musical communication through expressive use of timbre: The performer's perspective. *Psychology of Music*, 40(3), 301-323.

Horne, M. A. (1988). Towards a quantified, focus-based model for synthesizing English sentence intonation. *Lingua,* 75, 25 - 54.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, Massachusetts: MIT Press.

Huron, D. (2011). Why is sad music pleasurable? A possible role for prolactin. *Musicae Scientiae,* 15, 146–158.

Huron, D., and Royal, M. (1996). What is melodic accent? Converging evidence from musical practice. *Music Perception*, 13, 489–516.

Huttenlocher, P. (2002). *Neural plasticity*. Cambridge, MA: Harvard University Press.

Iakimova, G., Passerieux, C., Foynard, M., Fiori, N., Besche, C., Laurent, J.P., and Hardy-Baylé,M.C. (2009). Behavioral measures and event-related potentials reveal different aspects of sentence processing and comprehension in patients with major depression. *Journal of Affective Disorders, 113*, 188–194.

Ilie, G., and Thompson, W. F. (2011). Experiential and cognitive changes following seven minutes exposure to music and speech. *Music Perception,* 28, 247–264.

Ito, T., Murano, E. Z., and Gomi, H. (2004). Fast force-generation dynamics of human articulatory muscles. *Journal of Applied Physiology,* 96(6), 2318–2324.

Jiang, A., Yang, J., and Yang, Y. (2014). MMN responses during implicit processing of changes in emotional prosody: an ERP study using Chinese pseudo-syllables. *Cognitive Neurodynamics*, 8(6), 499–508.

Jones, M. R. (1987). Dynamic pattern structure in music: Recent theory and research. *Perception & Psychophysics*, 41, 621–634.

Jones, M. R. (1993). Dynamics of musical patterns: How do melody and rhythm fit together? In T. J. Tighe, and W. J. Dowling (Eds.), *Psychology and music: The understanding of melody and rhythm* (pp. 67–92). Hillsdale, NJ: Erlbaum.

Juslin, P. N. (1997). Perceived emotional expression in synthesized performances of a short melody: capturing the listener's judgment policy. *Musicae Scientiae*, 1, 225–256.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 1797–1813.

Juslin, P. N. (2001). Communicating emotion in music performance: A review and a theoretical framework. In P. N. Juslin, and J.A. Sloboda (Eds.), *Music and Emotion: Theory and Research* (pp. 309-337). New York: Oxford University Press.

Juslin, P. N. (2003). Communicating emotion in music performance: Review and theoretical framework. In P. N. Juslin and J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 309–337). Oxford: Oxford University Press.

Juslin, P. N., and Laukka, P. (2003). Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code? *Psychological Bulletin,* 129, 770–814.

Juslin, P. N. and Madison, G. (1999). The role of timing patterns in the decoding of emotional expressions in music performances. *Music Perception,* 17, 197-221.

Juslin, P. N., and Sloboda, J. A. (2013). Music and emotion. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 583–645). Amsterdam, the Netherlands: Elsevier.

Juslin, P. N., and Timmers, R. (2010). Expression and communication in music performance. In P. N. Juslin and J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 453–489). Oxford, UK: Oxford University Press.

Kaganovich, N., Kim, J., Herring, C., Schumaker, J., MacPherson, M., and Weber–Fox, C. (2013). Musicians show general enhancement of complex sound encoding and better inhibition of irrelevant auditory change in music: an ERP study. *European Journal of Neuroscience*, 37, 1295–1307.

Kanske, P., and Kotz, S. A. (2007). Concreteness in emotional words: ERP evidence from a hemifield study. *Brain Research*, 1148, 138–148.

Kanske, P., Plitschka, J., and Kotz, S. A. (2011). Attentional orienting towards emotion: P2 and N400 ERP effects. *Neuropsychologia,* 49, 3121–3129.

Kaplan, A. (2010). *Phonology shaped by phonetics: The case of intervocalic lenition*. PhD dissertation. University of California, Santa Cruz.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., and Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modelling. *Journal of the Acoustical Society of America,* 77, 266–280.

Kienast, M., and Sendlmeier, W. F. (2000). Acoustical analysis of spectral and temporal changes in emotional speech. In *Proceedings of the ISCA ITRW on Speech and Emotion* (pp. 92–97). Belfast, Northern Ireland.

Kinoshita, H., Furuya, S., Aoki, T., and Altenmüller, E. (2007). Loudness control in pianists as exemplified in keystroke force measurements on different touches. *Journal of the Acoustical Society of America*, 121, 2959–2969.

Kirchner, R. (1998). *An effort-based approach to consonant lenition*. PhD dissertation. UCLA.

Kisilevsky, B.S., Hains, S. M. J., Lee, K., Xie, X., Huang, H., Ye, H.-H., Zang, K., and Wang, Z. (2003). Effects of experience on fetal voice recognition. *Psychological Science*, 14, 220–224.

Klauer, K., and Musch, J. (2003). Affective priming: Findings and theories. In J. Musch and K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 7–50). Mahwah, NJ: Erlbaum.

Kochevitsky, G. (1967). *The art of piano playing: A scientific approach.* Princeton, NJ: Summy-Birchard Music.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., and Friederici, A. D. (2004). Music, language and meaning: Brain signatures of semantic processing. *Nature Neuroscience*, 7, 302–307.

Koelsch, S., and Siebel, W. A. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences*, 9, 578–584.

Kogan, N. (1997). Reflections on aesthetics and evolution. *Critical Review*, 11, 193-210.

Kong, Y. Y., and Zeng, F. G. (2006). Temporal and spectral cues in Mandarin tone recognition. *Journal of the Acoustical Society of America*, 120, 2830 – 2840.

Kotz, S. A., and Paulmann, S. (2011). Emotion, language, and the brain. *Language and Linguistics Compass*, 5, 108–125.

Kreiman, J., and Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Wiley-Blackwell.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch.* New York: Oxford University Press.

Krumhansl, C. L. (1995a). Effects of musical context on similarity and expectancy. *Systematische Musikwissenschaft, 3*(2), 211–250.

Krumhansl, C. L. (1995b). Music psychology and music theory: Problems and prospects. *Music Theory Spectrum*, 17, 53–90.

Krumhansl, C. L. (2010). Plink: "Thin slices" of music. *Music Perception*, 27, 337–354.

Kuhl, P. K. (1988). Auditory perception and the evolution of speech. *Human Evolution*, 3, 19-43.

Kutas, M., and Hillyard, S. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207, 203–205.

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621–627.

Ladd, D. R. (2008). *Intonational phonology (2nd edition).* Cambridge University Press.

Ladd, D. R., and Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics*, 25, 313–342.

Lai, C. (2009). Perceiving surprise on cue words: Prosody and semantics interact on *right* and *really*. In *Proceedings of Interspeech* (pp.1963-1966), Brighton, UK.

Lartillot, O. (2014). MIRtoolbox user's guide 1.6.1. Retrieved from https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/MIRtoolbox1.6.1guide.

Latinus, M., and Belin, P. (2011). Human voice perception. *Current Biology*, 21, 143–145.

Laukka, P., Juslin, P.N., Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, 19 (5), 633–653.

Lavoie, L. M. (2001). *Consonant strength: Phonological patterns and phonetic manifestations*. New York, NY: Garland.

LeDoux, J. (1996). *The Emotional Brain*. New York, NY: Simon and Schuster.

Lerdahl, F. (2013). Musical syntax and its relation to linguistic syntax. In M. A. Arbib (Ed.), *Language, music, and the brain: A mysterious relationship* (pp. 257-272). Cambridge, MA: MIT Press.

Lerdahl, F., and Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Lerdahl, F., and Krumhansl, C. (2007). Modeling tonal tension. *Music Perception*, 24, 329- 366.

Levy, D. A., Granot, R., and Bentin, S. (2001). Processing specificity for human voice stimuli: Electrophysiological evidence. *NeuroReport*, 12, 2653–2657.

Levy, D. A., Granot, R., and Bentin, S. (2003). Neural sensitivity to human voices: ERP evidence of task and attentional influences. *Psychophysiology*, 40, 291–305.

Liberman, A. M., and Mattingly, I. G. (1989). A specialization for speech perception. *Science*, 243, 489–494.

Liberman, M. Y., and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In Mark M. Aronoff, and R. T. Oehrle (Eds.), *Language sound structure: Studies in phonology presented to Morris Halle* (pp.157–233). Cambridge, MA: MIT Press.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773-1781.

Lindblom, B. (1983). Economy of speech gestures. In P. MacNeilage (Ed.), *The production of speech* (pp. 217-245). New York, NY: Springer-Verlag.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. J. Hardcastle, and A. Marchal (Eds.), *Speech production and speech modelling* (pp. 413–415). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lindblom, B., and Sundberg, J. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, 50(4), 1166–1179.

Lindquist, K., Wager, T. D., Kober, H., Bliss-Moreau, E., and Barrett, L. F. (2012). The brain basis of emotion: A meta analytic review. *Behavioral and Brain Sciences*, 35, 121– 143.

Liu, F., and Xu, Y. (2005). Parallel Encoding of Focus and Interrogative Meaning in Mandarin Intonation. *Phonetica*, 62, 70 –87.

Livingstone, S. R., Thompson, W. F., Wanderley, M. M., and Palmer, C. (2015). Common cues to emotion in the dynamic facial expressions of speech and song. *The Quarterly Journal of Experimental Psychology*, 68, 952-970.

Lopez-Calderon, J., and Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers of Human Neuroscience*, 8, 213.

Loucks, T. M., Ofori, E., Grindrod, C. M., De Nil, L.F., and Sosnoff, J. J. (2010). Auditory motor integration in oral and manual effectors. *Journal of Motor Behaviour*, 42, 233–239.

Madison, G. (2000). Properties of expressive variability patterns in music performances. *Journal of New Music Research*, 29, 335-356.

Malécot, A. (1955). An experimental study of force of articulation. *Studia Linguistica*, 9, 35–44.

McAdams, S., Beauchamp, J., and Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *Journal of the Acoustical Society of America*, 105, 882-897.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psycholollgical Research*, 58, 177–192.

McPherson, A. (2013). Portable measurement and mapping of continuous piano gesture. In *Proceedings of the 13th International Conference on New Interfaces for Musical Expression (NIME)*, Seoul, South Korea.

McPherson, A., and Kim, Y. (2011). Multidimensional gesture sensing at the piano keyboard. In *Proceedings of the 29th ACM Conference on Human Factors in Computing Systems (CHI)* (pp.2789-2798). Atlanta, GA.

McPherson, A., and Kim, Y. (2013). Piano technique as a case study in expressive gestural interaction. In S. Holland, K. Wilkie, P. Mulholland, and A. Seago (Eds.), *Music and Human-Computer Interaction* (pp. 123–138). London: Springer.

McPherson, W. B., and Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology*, 36, 53–65.

Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.

Meyer, M., Baumann, S., and Jancke, L. (2006). Electrical brain imaging reveals spatio-temporal dynamics of timbre perception in humans. *NeuroImage*, 32, 1510–1523.

Minetti, A. E., Ardigo, L. P., and McKee, T. (2007). Keystroke dynamics and timing: accuracy, precision and difference between hands in pianist's performance. *Journal of Biomechanics*, 40, 3738–3743.

Monahan, C. B., Kendall, R. A., and Carterette, E. C. (1987).The effect of melodic and temporal contour on recognition memory for pitch change. *Perception and Psychophysics*, 41(6), 576–600.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist,* 111, 855-869.

Munhall, K. G., Ostry, D. J., and Parush, A. (1985). Characteristics of velocity profiles of speech movements. *Journal of Experimental Psychology*, 11, 457–474.

Münte, T. F., Altenmüller, E., and Jäncke, L. (2002). The musician's brain as a model of neuroplasticity. *Nature Reviews*, 3, 473–478.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realisation model*. Chicago: University of Chicago Press.

Narmour, E. (1991a). The top-down and bottom-up systems of musical implication: Building on Meyer's theory of emotional syntax. *Music Perception*, 9, 1-26.

Narmour, E. (1991b). The melodic structures of music and speech: Applications and dimensions of the implication-realization model. In J. Sundberg, L. Nord, and R. Carlson (Eds.), *Music, language, speech and brain* (pp.48-56). London: MacMillan Academic and Professional Ltd.

Narmour, E. (1992). *The analysis and cognition of melodic complexity: The implication-realisation model*. Chicago: University of Chicago Press.

Nelson, W. L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics*, 46, 135–147.

Neuhaus, H. (1973). *The art of piano playing*. New York: Praeger Publishers.

Noble, L., and Xu, Y. (2011). Friendly Speech and happy speech – Are they the same? In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp.1502-1505). Hong Kong.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, 41, 1-16.

Olson, H. F. (2003). *Music, physics and engineering* (2nd ed.). New York: Dover.

Opolko, F., and Wapnick, J. (2006). *The McGill University Master Samples Collection on DVD (3 DVDs)*. Quebec, Canada: McGill University.

Orgs, G., Lange, K., Dombrowski, J.-H., and Heil, M. (2006). Conceptual priming for environmental sounds and words: An ERP study. *Brain and Cognition*, 62, 267–272.

Orgs, G., Lange, K., Dombrowski, J.-H., and Heil, M. (2007). Is conceptual priming for environmental sounds obligatory? *International Journal of Psychophysiology*, 65, 162–166.

Orgs, G., Lange, K., Dombrowski, J.-H., and Heil, M. (2008). N400-effects to task-irrelevant environmental sounds: Further evidence for obligatory conceptual processing. *Neuroscience Letters*, 436, 133–137.

Ostry, D., Keller, E., and Parush, A. (1983). Similarities in the control of speech articulators and the limbs: Kinematics of tongue dorsum movement in speech. *Journal of Experimental Psychology*, 9, 622-636.

Ostry, D. J., and Munhall, K. G.(1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America*, 77, 640–648.

Padgett, J. (2009). Systemic contrast and Catalan rhotics. *The Linguistic Review*, 26, 431–463.

Paeschke, A., Kienast, M., and Sendlmeier, W. F. (1999). F0-contours in emotional speech. In *Proceedings of the International Congress of Phonetic Sciences* (pp.929-932). San Francisco, CA.

Painter, J. G., and Koelsch, S. (2011). Can out-of-context musical sounds convey meaning? An ERP study on the processing of meaning in music. *Psychophysiology*, 48, 645–655.

Palmer, C. (2006). The nature of memory for music performance skills. In E. Altenmüller, M. Wiesendanger, and J. Kesselring (Eds.), *Music, motor control and the brain* (pp. 39-53). Oxford, UK: Oxford University Press.

Palmer, C., Carter, C., Koopmans, E., and Loehr, J. D. (2007). Movement, planning, and music: Motion coordinates of skilled performance. In *Proceedings of the International Conference on Music Communication Science* (pp.119-122). Sydney, Australia.

Palmer, C., Koopmans, E., Loehr, J., and Carter, C. (2009). Movement-related feedback and temporal accuracy in clarinet performance. *Music Perception*, 26(5), 439–450.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions.* New York, NY: Oxford University Press.

Panksepp, J., and Watt, D. (2011). What is basic about basic emotions? Lasting lessons from affective neuroscience. *Emotion Review*, 3, 387– 396.

Pantev, C., Oostenveld, R., Engelien, A., Ross, B., Roberts, L. E., and Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature*, 392, 811–814.

Parncutt, R. (1989). *Harmony: A psychoacoustical approach*. Berlin: Springer-Verlag.

Parncutt, R. (2003). Accents and expression in piano performance. In K. W. Niemöller (Ed.), *Perspektiven und Methoden einer Systemischen Musikwissenschaft* (pp.163-185). Frankfurt/Main, Germany: Peter Lang.

Parncutt, R., Sloboda, J. A, Clarke, E.F., Raekallio, M., and Desain, P. (1997). An ergonomic model of keyboard fingering for melodic fragments. *Music Perception*, 14, 341–382.

Patel, A. D. (2008). *Music, language and the brain*. Oxford: Oxford University Press.

Patel, A. D., Iverson, J. R., and Rosenberg, J. D. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *Journal of the Acoustical Society of America*, 119, 3034 –3047.

Patel, A. D., and Peretz, I. (1997). Is music autonomous from language? A neuropsychological appraisal. In I. Deliege and J. A. Sloboda (Eds.), *Perception and cognition of music* (pp. 191– 215). Hove, England: Erlbaum.

Patel, A. D., Xu, Y., and Wang, B. (2010). The role of F0 variation in the intelligibility of Mandarin sentences. In *Proceedings of Speech Prosody 2010* (paper 890). Chicago, IL.

Paulmann, S., Bleichner, M., and Kotz, S. (2013). Valence, arousal, and task effects in emotional prosody processing. *Frontiers in Psychology*, 4, 345.

Paulmann, S., and Kotz, S. A. (2008). Early emotional prosody perception based on different speaker voices. *Neuroreport,* 19, 209–213.

Paulmann, S., Ott, D.V., and Kotz, S. A. (2011).Emotional speech perception unfolding in time: the role of the basal ganglia. *PLoS ONE,* 6:e17694.

Paulmann, S., and Pell, M. D. (2010). Contextual influences of emotional speech prosody on face processing: how much is enough? *Cognitive Affective and Behavioral Neuroscience*, 10(2), 230242.

Pell, M. D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S., and Rigoulot, S. (2015). Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biological Psychology*, 111, 14-25.

Peretz, I. (2006). The nature of music from a biological perspective. *Cognition*, 100, 1-32.

Peretz, I. (2012). Music, lanugage, and modularity in action. In P. Rebuschat, M. Rohrmeier, J. A. Hawkins, and I. Cross (Eds.), *Language and music as cognitive systems* (pp. 254–268). Oxford: Oxford University Press.

Peretz, I., Gagnon, L., and Bouchard, B. (1998). Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. *Cognition*, 68, 111–141.

Peretz, I., and Hyde, K. (2003). What is specific to music processing? Insights from congenital amusia. *Trends in Cognitive Sciences*, 7(8), 362–367.

Perkell, J. S., Zandipour, M., Matthies, M. L., and Lane, H. (2002). Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *Journal of the Acoustical Society of America*, 112 (4), 1627–1641.

Phan, K. L., Wager, T., Taylor, S.F., and Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage*, 16, 331– 348.

Pierrehumbert, J., and Beckman, M. (1988). *Japanese tone structure*. Cambridge, MA: The MIT Press.

Pinker, S. (2007). Toward a consilent study of literature. In: J. Gottschall and D.S. Wilson (Eds.), *The literary animal: Evolution and the nature of narrative.* Evanston: Northwestern Univ. Press.

Prom-on, S., Xu, Y., and Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, 125, 405-424.

Pulvermüller, F., and Shtyrov, Y. (2006). Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes. *Progress in Neurobiology*, 79, 49–71.

Rainville, P., Bechara, A., Naqvi, N., and Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, 61, 5–18.

Reisenzein, R. (2000). Exploring the Strength of Association Between the Components of Emotion Syndromes: The Case of Surprise. *Cognition and Emotion*, 14, 1–38.

Repp, B. H. (1992a). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei". *Journal of the Acoustical Society of America,* 92*,* 2546-2568.

Repp, B. H. (1992b). A constraint on the expressive timing of a melodic gesture: Evidence from performance and aesthetic judgment. *Music Perception,* 10*,* 221-242.

Repp, B. H. (1994a). Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study. *Psychological Research,* 56*,* 269-284.

Repp, B. H. (1994b). On determining the basic tempo of an expressive music performance. *Psychology of Music*, 22, 157-167.

Repp, B. H. (1995). Expressive timing in Schumann's "Träumerei": An analysis of performances by graduate student pianists. *Journal of the Acoustical Society of America,* 98*,* 2413-2427.

Repp, B. H. (1996). The dynamics of expressive piano performance: Schumann's "Träumerei" revisited. *Journal of the Acoustical Society of America,* 100*,* 641-650.

Rietveld, A.C.M., and Gussenhoven, C. (1985). On the relation between pitch excursion size and pitch prominence. *Journal of Phonetics* 15, 273-285.

Rigoulot, S., Pell, M. D., and Armony, J. L. (2015). Time course of the influence of musical expertise on the processing of vocal and musical sounds. *Neuroscience*, 290, 175-184.

Robinson, K., and Patterson, R. D. (1995). The duration required to identify the instrument, the octave, or the pitch chroma of a musical note. *Music Perception*, 13, 1–15.

Rogier, O., Roux, S., Belin, P., Bonnet-Brilhault, F., and Bruneau, N. (2010). An electrophysiological correlate of voice processing in 4- to 5-year old children. *International Journal of Psychophysiology*, 75, 44–47.

Rossano, F., Carpenter, M. and Tomasello, M. (2012). One-year-old infants follow others' voice direction. *Psychological Science*, 23(11), 1298-1302.

Rozenkrants, B., Olofsson, J. K., and Polich, J. (2008). Affective visual event-related potentials: Arousal, valence, and repetition effects for normal and distorted pictures. *International Journal of Psychophysiology*, 67, 114–123.

Rump, H. H., and Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech*, 39, 1–17.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161–1178.

Sandor, G. (1981). *On piano playing: Motion, sound and expression*. New York: Schirmer Books.

Sauter, D. A., and Eimer, M. (2010). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience*, 22, 474 – 481.

Schellenberg, E. G., Iverson, P., and McKinnon, M.C. (1999). Name that tune: Identifying popular recordings from brief excerpts. *Psychonomic Bulletin and Review*, 6, 641–646.

Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In C. E. Izard (Ed.), *Emotions in personality and psychopathology* (pp. 493-529). New York: Plenum.

Scherer, K. R. (1989). Vocal correlates of emotion. In H. Wagner, and A. Manstead (Eds.), *Handbook of psychophysiology: Emotion and social behavior* (pp.165-197). London: Wiley.

Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice*, 9, 235–248.

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

Scherer, K. R., and Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1, 331-346.

Scherer, K. R., Zentner, M. R., and Stern, D. (2004). Beyond Surprise: The Puzzle of Infants' Expressive Reactions to Expectancy Violation. *Emotion*, 4, 389–402.

Schirmer, A., Chen, C., Ching, A., Tan, L., and Hong, R. (2013). Vocal emotions influence verbal memory: neural correlates and interindividual differences. *Cognitive, Affective and Behavioral Neuroscience*, 13, 80-93.

Schirmer, A., and Kotz, S. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, 10, 25-30.

Schirmer, A., Kotz, S., and Friederici, A. D. (2002). Sex differentiates the role of emotional prosody during word processing. *Cognitive Brain Research*, 14, 228-233.

Schwartz, G. E., Weinberger, D. A., and Singer, J. A. (1981). Cardiovascular differentiation of happiness, sadness, anger, and fear following imagery and exercise. *Psychosomatic Medicine*, 43, 343–364.

Seifert, U., Verschure, P., Arbib, M., Cohen, A., Fogassi, L., Fritz, T., Kuperberg, G., Manzolli, J., and Rickard, N. (2013). Semantics of internal and external worlds. In M. A. Arbib (Ed.), *Language, music, and the brain: A mysterious relationship* (pp. 203-232). Cambridge, MA: MIT Press.

Seppi, D., Batliner, A., Steidl, S., Schuller, B., and Nöth, E. (2010). Word Accent and Emotion. In *Proceedings of Speech Prosody*. Chicago, IL.

Shaver, P. R., Schwartz, J., Kirson, D., and O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52, 1061–1086.

Sityaev, D., and House, J. (2003). Phonetic and phonological correlates of broad, narrow and contrastive focus in English. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1819-1822). Barcelona.

Slevc, L. R. (2012). Language and music: sound, structure, and meaning. *Cognitive Science*, 3(4), 483-492.

Sloboda, J., and O'Neill, S. (2001). Emotions in everyday listening to music. In P. Juslin and J. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 415–429). Oxford, England: Oxford University Press.

Spencer, H. (1857). *The origin and function of music*. Fraser's Magazine, 56, 396–408.

Spreckelmeyer, K. N., Altenmüller, E., Colonius, H., Münte, T. F. (2013). Preattentive processing of emotional musical tones: a multidimensional scaling and ERP study. *Frontiers in Psychology*, 4, 656.

Spreckelmeyer, K. N., Kutas, M., Urbach, T. P., Altenmuller, E., and Munte, T. F. (2006). Combined perception of emotion in pictures and musical sounds. *Brain Research*, 1070 (1), 160–170.

Stein, R. B. (1982). What muscle variables does the central nervous system control? *Behavioural and Brain Sciences*, 5, 535-578.

Steinbeis, N., and Koelsch, S. (2008). Shared neural resources between music and language indicate semantic processing of musical tension resolution patterns. *Cerebral Cortex*, 18, 1169 –1178.

Steinbeis, N., and Koelsch, S. (2011). Affective priming effects of musical sounds on the processing of word meaning. *Journal of Cognitive Neuroscience*, 23, 604–621.

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: The MIT Press.

Sundberg, J. (1982). Speech, song, and emotions. In M. Clynes (Ed.), *Music, mind, and brain: The neuropsychology of music* (pp. 137–149). New York: Plenum Press.

Sundberg, J. (2000). Four years of research on music and motion. *Journal of New Music Research*, 29, 183-185.

Terken, J. M. B., and Hermes, D. J. (2000). The perception of prosodic prominence. In M. Horne (Ed.), *Prosody: Theory and experiment. Studies presented to Gösta Bruce* (pp. 89-127). Dordrecht: Kluwer.

't Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America*, 67, 811–821

't Hart, J., Collier, R., and Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Thomassen, J. M. (1982). Melodic accent: Experiments and a tentative model. *Journal of the Acoustical Society of America,* 71, 1596–1605.

Thompson, W. F., Cuddy, L. L., and Plaus, C. (1997). Expectancies generated by melodic intervals: Evaluation of principles of melodic implication in a melody-completion task. *Perception and Psychophysics*, 59(7), 1069–1076.

Tortora, G. J. (2002). *Principles of human anatomy*. New York, NY: Wiley.

Traunmüller, H. and Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, 107(6), 3438-3451.

Trehub, S., Endman, M. W., and Thorpe, L. A. (1990). Infants' perception of timbre: Classification of complex tones by spectral structure. *Journal of Experimental Child Psychology*, 49(2), 300–313.

Uchanski, R. M. (2008). Clear speech. In D. B. Pisoni, and R. E. Remez (Eds.), *The handbook of speech perception* (pp. 207–235). Oxford, UK: Blackwell.

Vaish, A., Grossmann, T., and Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, 134, 383–403.

van Petten, C. and Rheinfelder H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, 33, 485-508.

van Vugt, F. T., Furuya, S., Vauth, H., Jabusch, H.-C., Altenmüller, E. (2014). Playing beautifully when you have to be fast: spatial and temporal symmetries of movement patterns in skilled piano performance at different tempi. *Experimental Brain Research*, 232 (11), 3555–3567.

Ververidis, D., and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48, 1162-1181.

Watson, D. (1991). *The Wordsworth dictionary of musical quotations*. Ware, England: Wordsworth.

Westbury, J. R., and Keating, P. A. (1986). On the naturalness of stop consonant voicing. *Journal of Linguistics*, 22(1), 145–166.

Widmer, G., and Goebl, W. (2004). Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, 33(3), 203–216.

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., and Ackermann, H. (2005). Identification of emotional intonation evaluated by fMRI. *NeuroImage*, 24, 1233–1241.

Williams, P. L., and Warwick, R. (1980). *Gray's anatomy*. Edinburgh: Churchill Livingstone.

Wilson, G. D. (1994). *Psychology for performing artists: Butterflies and bouquets.* London: Jessica Kingsley.
Winges, S. A., Furuya, S., Faber N. J., and Flanders, M. (2013). Patterns of muscle activity for digital coarticulation. *Journal of Neurophysiology*, 110, 230-242.

Xu, Y. (1999). Effect of tone and focus on the formation and alignment of F_0 contours. *Journal of Phonetics*, 27, 55–107.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46, 220-251.

Xu, Y. (2011). Post-focus compression: Cross-linguistic distribution and historical origin. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp.152-155). Hong Kong.

Xu, Y. (2014). *ProsodyPro.praat*. University College London, London, UK.

Xu, Y., Kelly, A. and Smillie, C. (2013a). Emotional expressions as communicative signals. In S. Hancil, and D. Hirst (Eds.), *Prosody and iconicity* (pp. 33-60). Amsterdam, The Netherlands: John Benjamins Publishing Company.

Xu, Y., Lee, A., Wu, W.-L., Liu, X., and Birkholz, P. (2013b). Human vocal attractiveness as signalled by body size projection. *PLoS ONE*, 8(4), e62397.

Xu, Y., and Prom-on, S. (2010-2015). *PENTAtrainer1.praat.* Retrieved from http://www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer1/

Xu, Y., and Wang, M. (2009). Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics*, 37, 502–520.

Zachar, P., and Ellis, R.D. (2012). *Categorical Versus Dimensional Models of Affect: A Seminar on the Theories of Panksepp and Russell. Consciousness and Emotion.* Amsterdam, the Netherlands: John Benjamins Publishing Company.

Zanon, P., and De Poli, G. (2003a). Estimation of parameters in rule systems for expressive rendering in musical performance. *Computer Music Journal*, 27, 29–46.

Zanon, P., and De Poli, G. (2003b). Time-varying estimation of parameters in rule systems for music performance. *Journal of New Music Research*, 32, 295–315.

Zatorre, R. (2013). Predispositions and plasticity in music and speech learning: neural correlates and implications. *Science*, 342, 585–589.

Zatorre, R. J., and Baum, S. R. (2012). Musical melody and speech intonation: Singing a different melody? *PLoS Biology*, 10, e1001372

Zhang, D., Liu, Y., Hou, X., Sun, G., Cheng, Y., and Luo, Y. (2014). Discrimination of fearful and angry emotional voices in sleeping human neonates: a study of the mismatch brain responses. *Frontiers in Behavioral Neuroscience*, 8, 422.