One music? Two musics? How many musics?

Cognitive ethnomusicological, behavioral, and fMRI study on vocal and instrumental rhythm processing.

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Hung, Tsun-Hui

Graduate Program in Music

The Ohio State University

2011

Dissertation Committee:

Udo Will, Advisor

Arved Ashby

Marjorie K.M. Chan
Abstract

The origins of music have been and still are a mystery; there is simply no good explanation or convincing evidence on how music actually started, yet. Fitch (2006) reviews a variety of studies on animal communication and indicates that unlike humans, animals usually produce either vocal or non-vocal sounds only. His research led to the suggestion that, in order to explore the origins of music it might be helpful not to treat music as a unitary phenomenon, but to distinguish clearly between vocally and non-vocally produced sounds and their usage. These ideas connect with comparative musicological studies showing, that even though every culture in the world has something we can call ‘music,’ many societies show a distinct preference for vocal music. Furthermore, numerous physiological and imaging studies have shown that speech and non-speech sounds are processed differently by the human brain (e.g., Belin et al., 2000). It has also been shown that vocal and instrumental melodic contours have differential effects on speech processing (Poss et. al, 2008). This dissertation explores whether there are also differences in processing of vocal and instrumental rhythms. A set of experiments investigates how humans process these two types of rhythm through reaction time measurements in behavioral experiments and through brain activation measurements in functional magnetic resonance imaging (fMRI) experiments. In each
trial two sound files, containing both vocal and instrumental rhythms were played, and participants had to decide whether either the vocal or the instrumental rhythms were the same. Results show that decisions on vocal rhythms caused higher brain activation than those on instrumental rhythms; these differences are mainly seen in early auditory pathways, e.g. temporal lobe, while activations for late auditory processing were similar for both rhythms. In the behavioral experiments participants show significantly shorter reaction time for vocal than for clapstick rhythms if rhythm pairs are the same. The results are in line with previous studies showing that vocal sounds and pitch contours are processed differently than instrumental or non-vocal ones. The fact that not only vocal and instrumental melodic contours but also vocal and instrumental rhythms are processed differently strongly supports the ideas about different phylogenetic origins of instrumental and vocal music. A preference to vocal music that is indicated by comparative musicological studies. In addition there is a preferred treatment of vocal sounds as demonstrated by many cognitive studies. This in turn seems to be a reflection of the importance of vocal communication in human societies. Our results are compatible with the idea that both vocal music and speech may originate from a common vocal communication system, whereas instrumental music originated from an accompanying system that produced sounds via manual activities.
Dedication

to my family
Acknowledgments

Thanks to my mentor Dr. Udo Will, for his advising throughout my Ph.D. study and this dissertation.

Thanks to Dr. Amir Abduljalil, for his encouraging advice, ample assistance, and full support during my fMRI data collecting, during the day, at night, over the weekends, and even on Christmas Eve.

Thanks to the Wright Center of Innovation, for making its excellent fMRI facility available for my study.

Thanks to all lovely experiment participants, for their time and ‘functioning brains.’

Thanks to Sishe Chin, for his beautiful graphic design, so my dissertation has more colors than just black and white.

Thanks to friends who brought food to me while I had no time to cook, so I did not need to experience too much American fast food.

Thanks to friends who exercised with me while I was writing, so I could exercise more than just my brain and fingers.

Thanks to the Musicology department at School of Music, for its continued financial support throughout my study at OSU.
Finally, many thanks to the Taoist Hsing Tian Kong (行天宮) scholarship for talented artists, for its continued blessing and support all the way through my undergraduate, master, and doctoral studies.

Therefore, I have made my dream…
Vita

2005 .................................................. B.F.A. Chinese music, Chinese Culture University, Taiwan.

2007 .................................................. M.M. Music Composition, Ohio University.

2007 to 2010 .................................... Graduate teaching and research associate, School of Music, The Ohio State University.

2008 .................................................. OSU Excavations at Isthmia, Greece.

2011 to present ................................. Lecturer, School of Music, The Ohio State University.

Publications


Fields of Study

Major Field: Music
Area of Emphasis: Cognitive Ethnomusicology
# Table of Contents

Abstract ............................................................................................................................................. ii

Dedication ........................................................................................................................................ iv

Acknowledgments ............................................................................................................................. v

Vita .................................................................................................................................................... vii

Publications ....................................................................................................................................... vii

Fields of Study ................................................................................................................................. vii

Table of Contents ............................................................................................................................. viii

List of Tables ..................................................................................................................................... xiii

List of Figures .................................................................................................................................. xiv

1 Origin and evolution of music diversity ......................................................................................... 1

1.1 Introduction ................................................................................................................................. 1

1.2 Documentations of early music--- some examples from Chinese Music ......................... 3

1.3 Examples from world music practices ..................................................................................... 4

1.4 Hints from biology and evolution of instrumental and vocal music .............................. 5

2 Music and Language ..................................................................................................................... 10

2.1 Introduction ................................................................................................................................. 10

2.2 Common topics for music and language studies: pitch and structure ....................... 11

2.3 Rhythm in music and language studies ............................................................................... 13

2.4 Speech surrogate--- a bridge between language and music ......................................... 14

viii
5.3.2 Participants........................................................................................................................................52
5.3.3 Experiment procedure.......................................................................................................................52

6 Behavior experiments and results ............................................................................................................54
6.1 Reaction time (RT) measurements.......................................................................................................54
6.1.1 Reaction time and participants’ response strategies ........................................................................56
6.1.2 Stimulus classification into small-integer and complex-integer ratio pattern 59

6.2 Participants............................................................................................................................................61
6.3 Task procedure and design....................................................................................................................61
6.4 Analysis................................................................................................................................................63
6.4.1 Error analysis....................................................................................................................................63
6.4.2 Reaction Time (RT)..........................................................................................................................65
6.4.3 The on-beat/off-beat effect ..............................................................................................................68
6.5 Discussion of errors...............................................................................................................................73
6.6 Discussion of the reaction time measures............................................................................................75
6.7 Summary of behavioral experiments....................................................................................................80

7 fMRI passive listening experiment .........................................................................................................82
7.1 Introduction to fMRI experiments.......................................................................................................82
7.1.1 Use of Scans......................................................................................................................................84
7.1.2 Participants.......................................................................................................................................87
7.1.3 Procedure .........................................................................................................................................87
7.1.4 Stimuli...............................................................................................................................................88
7.1.5 Scanning sequence............................................................................................................................88
7.1.6 fMRI data acquisition ....................................................................................................................88
7.1.7 Individual and group activation maps ................................................................. 90
7.2 Stimulus sequence ................................................................................................. 91
7.3 Task procedure ....................................................................................................... 92
7.4 Data analysis .......................................................................................................... 92
7.5 Results ..................................................................................................................... 93
  7.5.1 Stimulation with clapstick rhythms ................................................................. 93
  7.5.2 Stimulation with vocal rhythms ...................................................................... 94
  7.5.3 Comparison between two stimuli .................................................................... 96
  7.5.4 Stimuli of combined sounds ......................................................................... 99
  7.5.5 Stimuli of clapstick (the pulses of the two stimuli aligned) ....................... 100
8 fMRI decision task .................................................................................................. 106
  8.1 Stimulus sequence ............................................................................................... 106
  8.2 Task procedure .................................................................................................... 107
  8.3 Data analysis ........................................................................................................ 110
  8.4 Discussion ............................................................................................................ 121
9 General discussion .................................................................................................. 124
  9.1 Summary and discussion of experiments ......................................................... 125
  9.2 Music archaeology ............................................................................................. 129
  9.3 Physiological connections ................................................................................. 130
  9.4 Use of music and language .............................................................................. 132
  9.5 World music examples ...................................................................................... 133
10 Impact and future studies ...................................................................................... 137
  10.1 Implications for music theory and musicology ............................................... 137
10.2 Suggested future research................................................................. 138
10.3 New research perspective to ethnomusicology .............................. 139
Reference ............................................................................................... 141
List of Tables

Table 1: Error frequencies for the four conditions .......................................................... 64
Table 2: Activation areas for all stimuli in passive fMRI scans................................. 102
Table 3: Activation areas, z-value and coordinates of both stimuli .............................. 114
Table 4: T-tests and significances of the differences in % signal change between the vocal and the clapstick rhythm task. ................................................................. 120
List of Figures

Figure 1: Example from Silbo whistle language ..............................................15
Figure 2: Structure and frequency map of basilar membrane............................22
Figure 3: Map of cortical tonotopy ....................................................................26
Figure 4: Location of When, What, and Where pathways ...............................28
Figure 5: Distribution of sequential time interval ratios for clapstick and vocal rhythms in the 10 musical excerpts .................................................................46
Figure 6: One trial presentation ........................................................................49
Figure 7: Waveform of one stimulus trial ..........................................................50
Figure 8: Accepted reaction time range ..............................................................56
Figure 9: ANOVA of the reaction time with difference-point classes as factors ....58
Figure 10: The distribution of points of difference in the stimuli .....................59
Figure 11: Presentation sequence within one stimuli pair ..................................63
Figure 12: Mean % correct responses across subjects for stimulus conditions and clapstick and voice task .................................................................65
Figure 13: Interaction graph for stimulus conditions and task condition ..........66
Figure 14: Interaction graph for task condition and task-relevant stimulus condition ...67
Figure 15: Distribution of point-of -difference for vocal and clapstick rhythms in the same condition .................................................................68
Figure 16: Mean reaction time for the on- and off-beat condition ..................69
Figure 17: Mean reaction time for the on- and off-beat condition ....................71
Figure 18: Mean % correct responses in the on- and off-beat condition for clapstick and voice task. ........................................................................................................................................................................72

Figure 19: Mean correct responses across subjects for stimulus conditions, clapstick and voice task in experiment 2. ........................................................................................................................................................................................................................................73

Figure 20: Presentation sequence in the passive task .................................................................................................................................92

Figure 21: Average activations by clapstick stimuli. ........................................94

Figure 22: Average activations by vocal stimuli. ........................................95

Figure 23: Axial and coronal sections through the main activation areas in the passive listening task. .................................................................................................................................96

Figure 24: Axial and sagittal sections through highest activation areas. ..........98

Figure 25: Mean activation for vocal > clapstick. .......................................99

Figure 26: Activations of left and right hemispheres in the passive listening task........103

Figure 27: Stimulus presentation in fMRI decision task. ................................107

Figure 28: The task procedure.................................................................108

Figure 29: Stimulus sequence. ..................................................................109

Figure 30: Average activations of clapstick rhythm decision at different slice........111

Figure 31: Average activations of vocal rhythm decision at different slice. ........112

Figure 32: Coronal sections showing the main activation areas for clapstick and vocal rhythm decision task. .................................................................................................................................113

Figure 33: Coronal sections showing areas exclusively activated by the clapstick task and the vocal task.................................................................116

Figure 34: Coronal sections showing areas in which either the clapstick task or the vocal task produces higher activations. .................................................................................................................................117

Figure 35: Signal changes for clapstick and vocal rhythms tasks in the regions of interest (ROI). ........................................................................................................................................................................119
Figure 36: Location of Angular gyrus, supramarginal gyrus, and STG. ..........................120
Figure 37: Location of basal ganglia, putamen, caudate, and thalamus .........................121
Figure 38: Development of auditory system before and after birth..................................131
Figure 39: Development of manual dexterity and language..........................................132
1 Origin and evolution of music diversity

1.1 Introduction

Every human society has some kind of activities that we could call “music making.” Plato said “Music is a moral law. It gives soul to the universe, wings to the mind, flight to the imagination, a charm to sadness, gaiety and life to everything; It is the essence of order and lends to all that is good, just, and beautiful.” Music plays an important role in human life since time immemorial. It not only gives pleasure to humans as entertainment but also fulfills social functions in our daily life. To mention only a few: there are ritual and religious musics, work songs, and messages encoding melodies and rhythms. Without the practice of music, many societies might live in a very different way. Ethnomusicologists and musicologists have worked together closely to demonstrate and document how music is and was played, how it functioned, and how it is transmitted within and beyond a society. Therefore, we now understand that the practice of music seems to be universal but, at the same time, every society and culture has a distinctive identification and practice for their own music making. Thus, the world of music is always vivid as people continue to make music at every moment between sunrise and sunset.
Like the arts, where paintings can be presented by watercolor, oil painting, pastel, ink, acrylic, etc., music can be made with a variety of instruments, such as piano, shakuhachi, rebat, ud, djembe, etc, or by the human voice. Archaeological findings show that musical instruments might be in use since ancient times. For example bone flutes were found in Henan, China, have been dated to around 9000 years ago, at the beginning of Chinese civilization (early Neolithic). Tonal analysis of the flutes shows that they could already produce eight different pitches. The analysis suggests musicians from 7000 BC could have played not only single sounds, but also melodies on their instruments (Zhang et al., 1999). Another excavation in southwestern Germany found bone flutes that were even dated back to 35000 years ago. Such a sophisticated instrument implies that a refined tool making technology must already have existed at that time (Conard et al, 2009). Even if we cannot know for sure how the instruments were used or practiced, or how they were functioned, there is little doubt that people used tools to produce musical instruments since early human history.

Both instruments and the human voice are able to produce and manipulate sound in similar ways. For example, they can produce series of varying pitches, with changing dynamics, rhythmic patterns, and timbral variations that in turn allow musicians to make sequences of organized sounds that we consider music. However, in what way are instrumental music and vocal music the same? Do humans sense and feel these two types of music in similar ways? Do people show different preferences for instrumental music and vocal music? Does the human brain treat instrumental and vocal music the
same way? What about the origin of human music? Did instrumental music and vocal music develop simultaneously and in similar ways?

Perhaps examples from world musics can give answers to these questions.

1.2 Documentation of early music--- some examples from Chinese Music

According to historical documents (Li, 2001), even though the literature and archaeological findings were still obscure to scholars, it seems instrumental music may play a very important role in ancient Chinese court music, especially the music played at the early court (from Xia (夏) Dynasty ca. 2070 BC to Zhou (周) Dynasty ca 256 BC). However, court music was directed by the court and served clear functions of controlling and influencing people and society. The purpose of court music was not just enjoyment but also for practicing the power of the court. Therefore, music at the Chinese court is probably not a good example for observing people’s preferences for instrumental music or vocal music in early times.

On the other hand, one of the earliest documented folk music genres in China was Xiang he ge (相和歌, harmonious song) during Han (漢) Dynasty (206BC - 220AD). “Xiang he ge is mainly vocal, but silk and bamboo [instruments] may accompany the singing accordingly. The singer also plays a rhythmic percussion instrument to mark accents and meter, in the Book of Music.” (相和，漢舊歌也；絃竹更相和，執節者歌。晉書，樂誌。) This genre, which started as a vocal solo genre,
indicated that vocal music was dominant and the instruments were initially accompanying the voice.

In a Chinese book, the *Tales of the New Era* (世說新語) from 5th AD, we read that “string is not as good as bamboo; bamboo is not as good as voice.” (丝不如竹，竹不如肉。) (Wang, 1981). The stories in this book describe the real life and thoughts of the upper class citizen. This particular quote shows that people in ancient Chinese society attributed higher values to the voice and vocal music than to wind and string instruments.

Throughout the early Chinese pre-imperial music history, Chinese music seems to have clear preferences for vocal music. We have too little information about folk music before the Song (宋) Dynasty (960 AD to 1270 AD), but as instrumental music already played an important role in ritual music since early Chinese history, the pure instrumental folk music could also played outside of ritual events.

1.3 Examples from world music practices

Besides Chinese music, similar observations can be made on North American indigenous music, Indian music, Australian Aboriginal music, etc. Vocal music presents and plays a prominent role in all human societies, even though instruments are also in use. However, very often instrumental music is to accompany vocal music and vocalists sing the main melodies. Examples from North American native music show there are only a few instruments in their performances and these instruments usually play steady
beats as background accompaniment. Shamans (medicine men) from central Asia, South Africa, North America, or South America use few melodic instruments, and they tend to imitate the sounds of nature and animal calls through their own voice.

In various cultures important instrumental music genres developed based on vocal music practices. For example, the gayaki ang style in north-India is based on vocal performance styles. Musicians create special techniques and try to ‘sing’ through their instruments like the human voice; this musical style takes aesthetics and expressiveness of vocal music as a guideline and developed special techniques to put them into practice. For example the possibility of pulling strings sideways to vary the tension allows the sitar to increase its flexibility to imitate the human voice.

These observations lead to another question: perhaps the difference between vocal and instrumental music is not just in the physical features of the sounds. Could it be that there is also a difference in the cognitive processing of these two types of music that is linked to the origins of human action and communication?

1.4 Hints from biology and evolution of instrumental and vocal music

Chase (2001) has shown that even goldfish can learn to distinguish Baroque music from blues. This suggests that the earliest jawed vertebrates may already have certain mechanisms that may be homologous to those present in human sound processing and may also play an important role in music perception. Then this mechanism would date back to about 500 million years ago in the history of vertebrates.
Another type of animal study, already discussed by Aristotle, has shown that birds have an ability to learn songs. Darwin suggested birdsong and human music are evolutionary analogs, not homologous. That would mean that, although birdsongs and human music both contain melodic contours, birdsongs have a different evolutionary origin than music. However, the studies on birdsongs and songs of other animals suggest that a common feature of these and human songs may be vocal learning (Riebel, 2003): this ability was already available in these ‘singing’ non-human species, and it may also be one of the foundations of human vocal music making.

Examining closer related species we find that besides vocal sound making, primates like gorillas and bonobos also show the ability to produce rhythmic sound sequences by using other parts of their bodies. This ability can be considered analogous to human instrumental music making. For example, bonobos are reported to be able to maintain steady drummed beats for at least 12 seconds (Fitch, 2006). Comparative ethnographic studies seem to suggest drumming may be one of the widespread and oldest forms of instrumental music making. Thus, human and primate manual and bi-manual sound production like drumming may possibly be homologous, not just analogous phenomena.

In the human world, without any direct historical traces and evidence that could be examined today, the origin of human vocal activity, whether singing or speaking, is still speculative. The development of the vocal folds allows human to produce sounds via the vocal tract. In particular, the descended larynx is a critical element for the
development of human speech (Laitman, 1988). However, in order to speak humans not only need a vocal tract but also the required cognitive capabilities. Usually humans know how to use the vocal tract at birth, learn how to mimic sounds in early life, and then finally learn to speak. Scholarly opinions vary on the evolutionary timeline of evolution of speech. Nevertheless, vocal communication probably developed as early as the homo habilis era, ca. 2.3 to 2.4 million years ago. Archaeological observations strongly suggest that communication was necessary for complex artifact constructions, transmission of knowledge and social activities. Therefore the ability of human vocal communication may have developed 2 to 2.5 million years ago.

Much evidence indicates that humans have significantly been using their hands since early human development. The fact that humans acquired the ability to use vocal-auditory channels to communicate leaves the body free for other activities (Hockett, 1960), including collecting food, making tools and playing instruments. The ability of using hands allow humans to produce instrumental sounds like clapping or drumming on the body or other objects – an ability that is probably at the origin of instrumental music. Thus, humans were capable of producing and combining both instrumental and vocal sounds. On the other hand, many animals only have the ability to produce either vocal or instrumental sounds, but not both simultaneously.

With this diverse origin of instrumental and vocal sound, it is likely that their derivatives, instrumental and vocal music, might not be treated by the same ‘processing channels’ in human cognition. Even though both instrumental and vocal music are
called music, humans may still listen and process them in different ways. Research (Bent et al., 2006, Levy et al., 2001, Vouloumanos, 2001, & Zatorre et al., 2002) has shown that humans show different cognitive processing of speech and non-speech sounds. This suggests that people process the human voice different than other sounds. The degree of differences is influenced, among others, by people’s language background.

In addition, my previous research (Poss et al., 2008) on speech priming by vocal and instrumental melodic contours also shows that humans process the two types of sounds differently. In those experiments participants had to repeat target syllables orally after listening to a prime followed by a target. Vocal and instrumental melodic contours were used as primes. The results show a significant priming effect (i.e. facilitation of the speech task). The degree of priming depends on whether the pitch contours matched between prime and target as well as the timbre of prime (i.e. instrumental sounds vs. vocal sounds). One of the relevant results that connect to the current study was that vocal primes led to greater facilitation than instrumental primes. The study reinforces the idea that humans may process these two types of sounds differently.

Besides timbre and pitch, rhythm is one of the ubiquitous elements in music; it can signify emotion and meaning in music (Meyer, 1956). This dissertation is concerned with differences between vocal and instrumental music. In particular, it investigates whether there are differences in vocal and instrumental rhythm processing. I hope the results will align with what is known about the special role of the human
voice in perception and cognition, will illuminate the relationship between vocal and instrumental music, and will add to a better understanding of why most cultures in the world vocal music usually is dominate where instrumental music is accompany. Finally, the explorations of this dissertation will hopefully contribute to a better and more diversified understanding of this unique human activity that we call music and its foundations.
2 Music and Language

2.1 Introduction

Not until recent years have researchers began to have clearer ideas about how people process sounds. Music and language are transmitted in the auditory-temporal domain and learned through exposure to local variants. They draw on many of the same general cognitive capacities, such as memory, fine voluntary motor control, imitation, and jointly intended action (Jeannerod, 2010). Physiological studies have shown that sound perception can be understood as a complex process that comprises the transformation of sound vibrations into electrical impulses in the Cochlea, the transmission of these neural activities via the thalamus to the auditory cortex where frequency, intensity, and sound quality are processed. The auditory cortex communicates with other parts of the cortex that process sounds in contexts of varying complexity. The brain can identify and attribute meaning to the perceived sounds no matter whether these sounds are language, music, noise, or other types of environmental sounds. Language processing seems very natural to most of us, but it is one of the most complicated cerebral activities we engage in human cognition. More than one area of the brain from both right and left hemispheres is involved in processing language.
2.2 Common topics for music and language studies: pitch and structure

Patel (2008) states, “Every human infant is born into a world with two distinct sound systems. The first is linguistic and the vowels, consonants, and pitch contrasts of the native language. The second is musical and includes the timbres and pitches of the culture’s music.” In addition, our brain is obviously also capable of processing environmental and any other sounds in the world. With the help of modern advanced research technology, we have opportunities to discover how the brain processes music and language, and to understand the differences and commonalities between them.

Music as well as language is organized sound in time. The topics about the relationship between music and language, such as music in language (e.g. speech prosody) or language in music (e.g. vocal music), have gathered great importance and interest in recent years. However, during the last decades studies of music cognition have shifted from music psychological analysis (e.g. Huron, 2006) or borrowing models from linguistics, to more experimental and scientific stances. For example, in 1983 Fred Lerdahl and Ray Jackendoff published ‘A Generative Theory of Tonal Music’. This theory by Fred Lerdahl and Ray Jackendoff (1983), proposed that language and music have different components and attributions: Language is built of words and syntax, expressing propositional meaning; music has conventionalized sound patterns and hierarchical structures, expressing patterns of tension and relaxation or abstract event structures. While the evidence from a range of disciplines on the status of differences and similarities between music and language is constantly changing our view, the question of what language and music comprise as cognitive systems, as well as the
question about their relationship within human cognition, remains. The linguistic model that can or should be applied to music has been the subject of much discussion and debate ever since. For example the application of the theory of generative grammar to music assumes that music is cognitively processed like language, a view that leaves much to debate. Since the last ten years, new experimental and theoretical research approaches to music cognition have given us new opportunities and challenges for the investigation of the music/language relationship.

Patel (2003) suggests that musical melodies and speech have a common underlying cognitive basis, as is clear from his statement that “language and music rely on a shared system for structural integration. A single system was argued to be involved in integrating incoming elements (words/ notes) into evolving structures (sentences/ musical sequences).” He hypothesized that “the difficulty of structural integration in language and music is influenced by the distance between the incoming element and the element it needs to connect to (in language), or between the element and the harmonic context (in music).” In addition, he also argues (2008) that there might be a sharing of neural resources for dealing with syntactic mechanisms in music and language. His view on music syntax is analogous to Lerdahl and Jackendoff’s musical grammar. He thinks that there might be a central sequencing device for both music and language.

However, he also mentioned, “Music typically is seen as having a periodicity of pitch and of duration that everyday speech doesn’t (setting aside metric linguistic forms like poetic verse).” However, this is not evident and indeed, the argument has been made that both music and language show some kind of periodicity of pitch as well as
rhythmic structures. Therefore, Brown (2009) criticized that “Patel is too accepting of the differences between music and speech, and that future work in this area should strive toward a joint set of principles that can account for music and speech as variations on a common phonological theme of generating pitch sequences in time.”

2.3 **Rhythm in music and language studies**

   It seems most studies focus on either melodic contours and processing or the structures between music and language. Certainly, these two topics are relevant and very interesting for discussion. However, rhythm is also an important element in spoken languages. For example, different vowels and consonants have different length. There are languages (e.g., Hindi, Arabic) where phoneme length is a distinctive feature and consequently it is very important for a speaker to differentiate the length of vowels and consonants. In addition, in tone languages, particularly in Mandarin, the length of syllable may depend on the tones. In normal speech, syllables with the fourth tone are usually shorter than syllables with other tones where syllables with the third tone (dipping tone) are usually longer than syllables with other tones, if the contour is complete (Lee and Hung, 2008).

   So, rhythm is an important component in language and music production.

   Many people tend to make an automatic link between music and language, and think music is just another form of language. Speech surrogates might be one of a few situations in which both musical and linguistic features are produced as well as
perceived at the same time. Studies of speech surrogates may give us some hints concerning the relationship of language and music.

2.4 Speech surrogate--- a bridge between language and music

Like Theodore Stern (1957) has pointed out, human speech is a kind of signal system that can convey messages through sounds transmission. Another type of signal system is speech surrogacy, which is defined as the use of instruments other than the human voice to mimic the way in which speech communicates messages. Such signaling systems have been realized in various ways, with whistling, with flutes, or with drums. Typically they can be transmitted over longer distances than regular speech. They can be perceived over a range of about one or two kilometers, on high mountains even up to five kilometers given there are no obstacles such as trees or rocks (Meyer, 2007).

Whistle languages generally work with minimal dynamic variations and high pitches. For example, Silbo Gomero is a whistle language that has been and is practiced by shepherds and herders on the island of La Gomera in the Canary Islands, to overcome the difficulties of communication over long distances in an environment full of ravines and valleys as it is the case on this island. The whistle language has been adapted to these conditions and provides an island-wide communication system. Silbo Gomero is derived from a non-tonal language, Spanish. It is more like ordinary lip-whistling. The following spectrograms (Fig. 1) show the relationship between spoken
and whistled language: it is the second formant of the spoken language that corresponds to the fundamental pitch (f0) of the whistling.

Figure 1: Example from Silbo whistle language. Spectrograms of the vowel sequence /i e a o u/, as spoken (a) and whistled (b) (Rialland, 2005).

Instrumental surrogates are another type of surrogate language. Usually they are performed on a small instrument or on tools, for example, the drum. Drum language can be found in many regions in Africa especially where tone languages are spoken. The feature of different level tones in many African languages allows the drummers to
encode messages on drums with either variable pitch (e.g. hourglass-shaped drums like the tama (luna) or with multiple fixed pitches, like e.g. slit-drums.

In the study of drum speech surrogacy Túndé Adégbólá (2003) shows that the drum speech of the Yorùbá in West Africa (Nigeria, Benin and Togo) uses tones and rhythms to mimic human speech. Adégbólá looked at the uniqueness of linguistic features of the Yorùbá language and the way Yorùbá people manage to turn the drum surrogate into a functional means of communication. He discovered that the Yorùbá drum language has a special drum vocabulary, constructed of unique two-sentence structures (called gangan) with tones and rhythms. For example, the length of vowels can determine the rhythm of drum languages, and the drummers can control the pitch by changing tension of the drum skin. Another type of drum language is not based on the features of the spoken language but is made of arbitrary patterns used for signaling specific meaning. Overall, pitch and rhythm are the most important elements in drum languages for encoding and transmitting messages.

An fMRI study by Carreiras et al (2005) shows that when Silbo whistlers perceive the Silbo whistle sounds, their main language areas of the brain are activated, whereas people without the knowledge Silbo process the whistling as a collection of sounds and the brain activities were registered in other areas (Carreira et al., 2005). The results show that the brain may process the same information in different regions of the brain according to our experiences with the sounds. If one recognizes whistling as a language, then language areas of the brain are the main processing areas, if it is not recognized as language, other areas get involved. So, pitch and rhythm information that
encode spoken language in whistling, can activate/facilitate language processing and the degree of facilitation depends on familiarity with the surrogate.

2.5 **The relevance of cognitive science in ethnomusicology**

All mentioned studies in this chapter show how scientific studies can contribute much to music cognition research. Cognitive sciences have produced many studies on the relationship between music and language. Unfortunately, most studies were and still are focused on Western art music and are done by non-musicians. Thus, the results may not be able to present the cognitive process of the human mind in other (non-western) cultures. In addition, scientists without proper music training may focus more on the science than the music. Scientists often treat music just as *sound* while musicians treat the music as much more than just *sound*. Perhaps musicians as well as (ethno-) musicologists could contribute musical experience and knowledge to these research topics and give more input from the side of music and add a musicological dimension.

Music is a cultural product of human development (Cross, 2001), it offers an opportunity for us to understand the meaning of culture through performance practice. Yet, how do sounds become expressive music to us?

A task of ethnomusicological fieldwork is to make an effort to understand the music from the perspective of the performers and audiences, and what they say about the music (Will, 1998), by interviewing people. Yet, to interview people may not uncover all information that we need, because musical expressions sometimes cannot be articulated in words, and they are not as universal as we may think. Musical concepts
often do not translate across cultures, and sometimes we barely find an appropriate term for music from other cultures. For example, Central Australian aboriginal singers demonstrate that the concept of “pitch” is different for them than for western musicians (Will, 1998). Australian musicians do not have verbal expressions for the ways in which they organize pitch in their music. So, the method of interviewing may be limited when we want to understand the cultural meaning of a people’s music. Our hearing and perception have been formed and influenced by our culture. If we cannot hear everything and performers cannot tell us everything, then the only thing we can rely on fully is (a detailed analysis of) the music itself.

Music is not just an innocuous series of sounds, but it is influenced by, and it influences, the mind and the body of both performers and audiences. The mind is significantly linked with and influenced by both our body as well as our social environment (Núnez and Freeman, 1999), so music cannot be understood without acknowledging its surrounding environment. If either one of the factors mind, body, or social environment changes, the other factors must be affected in turn. Take as an example every musician’s performance experience that to play in a practice room is always different than playing on stage. “Performers and listeners’ response to music do not occur spontaneously” (Becker, 2001), but rather, their responses represent a cognitive process of a network of thoughts and behaviors. Therefore, to recognize their cultural and emotional expression and reaction from the emic perspective is as important as recognizing them from the etic one, because “what is not assumed in one mode may become central in another mode” (Becker, 2001).
Not only what is present on the music surface is important but also what is concealed. It is important to try to uncover concealed structures through analysis of the musicians’ non-verbal behaviors. Data from these analyses can then lead to hypotheses that may even be tested in psychological experiments. The hypotheses and the results of the experiments would allow us to describe musicians’ mental activities that influence their behaviors. And, subsequently, we can then search for an explanation of how these mental factors play a role in musical practice and the conscious mind.

Music cognition is research into the characteristics of mental activities and how these are related to and influence our musical behaviors. The “soundless” data, produced through analyses and experiments, offers another means for us to understand music of people.
3 Introduction to the auditory system

3.1 Human auditory system: how sounds make sense to us

The auditory system is one of the five sensory systems of our body. It gives us not only a sense for the world of sounds, but is also an important component of our ability to communicate with others. Especially language communication is one of the most important faculties for human societies. For many species, hearing ability helps animals to look for food (hunting), stay safe (protecting), seek a mate (courtship), etc. In this system, the ear is the organ which receives sounds and transforms them into neural activity. Three parts are included in the ear structure: outer ear, middle ear, and inner ear.

3.1.1 Structure of ear and its functions

3.1.1.1 Outer ear (or external ear)

The outer ear includes the pinna and external auditory meatus. Though one can still receive sounds without pinna, the shape of the outer ear helps our brain determine the sound direction. In addition, it amplifies 30- to 100-fold sound pressure for frequencies up around 3 kHz. This is one of the reasons that explain why humans are very sensitive to this frequency range and the bandwidth of most human speech sound is
also around 3kHz. External auditory meatus runs into middle ear. It leads the vibrations to the tympanic cavity and also amplified sounds from 3 to 12 kHz.

3.1.1.2 Middle ear

A separation between outer ear and middle is the eardrum or tympanic membrane. The middle ear still contains sound information in wave form. Three bones in the middle ear, the malleus (hammer), incus (anvil) and stapes (stirrup), convert low pressure vibrations to high pressure vibrations at the oval window of the inner ear.

3.1.1.3 Inner ear

At the oval window the vibrations of the ossicles are transformed into fluid motions inside the cochlea: The cochlea contains liquid, endolymph and perilymph. There are three sections in the cochlea: scala vestibule, scala media, and scala tympani. The inner ear is one of the most important parts of our auditory system. Any small damage in the inner ear could create severe hearing loss.

3.1.1.4 Basilar membrane (part of inner ear)

The basilar membrane, which is equipped with hair cells and two liquid-filled tubes inside of cochlea are the two main structures of the inner ear. Because of its taper and varying stiffness the basilar membrane causes dispersion of sound frequencies: for specific frequencies the vibration in certain locations is larger than other locations (Fig. 2). Georg von Békésy has shown that high frequencies cause maximum vibrations at the basal end and low frequencies cause maximum vibrations at the apical end. In other
words, different frequencies lead to maximum vibrations at different places on the basilar membrane of the cochlea. The vibrations excite the hair cells. Békésy’s research has shown that different frequencies are dispersed before activating the neurons which connect the cochlea to the auditory pathway and brain.

Figure 2: Structure and frequency map of basilar membrane (Port, 2007).

At the basilar membrane, the motion of the liquid wave is converted into electric signal. On the basilar membrane, there are thousands of inner and outer hair cells which
form the organ of the corti. The deflection of the hair cells opens ion channels so that any positively charged ions can pass through the cells. The hair cells do not fire an action potential by themselves, but the ions depolarize the cells and cause a receptor potential. The neurotransmitters then activate action potentials in the secondary neurons that innervate the cochlea. This is the point where the sound signal is converted into electrical nerve signal. Each hair cell is usually surrounded by supporting cells. The function of supporting cells is to separate one hair cell from another as well as support and protect the hair cells. The spiral ganglia in the cochlea have axons from the eighth cranial nerve (the vestibulocochlear nerve) and allow sending a neural representation of sound to brain. The receptors in the cochlea are where the auditory pathway begins.

3.1.1.5 Medulla, brainstem, hindbrain, midbrain

From the cochlea, information is send to the medulla oblongata of the brainstem, (which includes dorsal cochlear nuclei and superior olive), the inferior colliculus (midbrain), and, finally, the auditory cortex.

Each dorsal nucleus of the trapezoid body, receiving information from the cochlear nuclei, detects phase and intensity. In addition, the colliculus contains neurons that are responsible for detecting phase and intensity difference between two ears. The colliculus is also part of a higher level reflex pathway that is activated by sudden loud sounds.

There are two cues for spatial localization: inter-aural intensity difference (IID) and inter-aural timing difference (ITD). IID is believed to be processed and mediated by
lateral superior olive and ITD is by medial superior olive. The brain is able to detect very small timing differences when sounds arrive first in one ear and then the other, as well as the sound intensities differences when ears receive sounds from different angles. This information helps our brain to extract information about the direction as well as distance of the sounds. Even though the timing and intensity difference are very small, these infinitesimal disparities are sufficient for spatial localization by our brain.

3.1.2 Auditory system in the brain

Studies on auditory processing in medulla, brainstem, hindbrain, and midbrain are available through imaging methods like EEG, MEG, PET and fMRI. However, the resolution is very different among the methods. For example there are some activities in the brain that can only be observed by using EEG but not MEG, and vice versa. There are also brain regions that are more difficult to explore than others. For example the brainstem is difficult to investigate with fMRI, but even employing EEG technologies require special procedures due to the minute signals that can be registered from it at the head (or neck) surface. Therefore, different methods need to be applied depending on the research questions.

How do we understand sounds, what are the sounds, and where do the sounds come from? Many different types of sounds may occur on one occasion and we usually are able to hear all of them. For example, on the street, we may hear people’s talking, noise of shoes from walking, noise of a car engine, or construction noise. How do we process all the sounds together at the same time, or perhaps sort out the important
sounds (information) and neglect those that are not relevant? The auditory cortex can process this important and complicate task.

3.1.2.1 Primary auditory cortex

The primary auditory cortex is at the sylvian sulcus of the temporal lobe and serves higher auditory processing. Clinical studies have shown that patients with lesions in the primary auditory cortex have lost their sensation of sound while brainstem and midbrain reflexes to auditory stimuli are still available. The lesion may disrupt the transition from primary auditory cortex to higher processing levels in the auditory cortex.

Tonotopic organization refers to the anatomical mapping between different parts of a nervous system, also known as the frequency map of the brain---due to the fact that different frequencies are processed in slightly different places of the brain. This is similar to retinotopy in the visual system. Neurons in the primary auditory cortex are arranged according to the sound frequency to which they respond best. Neurons at the distal side of primary auditory cortex respond to low frequencies and neurons at the proximal side respond to high frequencies (see fig. 3). The tonotopic organization is already formed when frequencies are transformed into neuronal responses at the basilar membrane in the cochlea. The cochlear nuclei and most subsequent projection areas are also arranged tonotopically, so there is a chain of tonotopical maps along the auditory pathways, from the cochlea up to the auditory cortex (Heeger, 2006).
After the primary auditory cortex, the secondary auditory cortex is known for more complex processing of the spectral and temporal information of auditory signals. For example, the Wernicke area is involved in understanding of spoken and written language.

Another important area for sound processing is the auditory association cortex. It surrounds the auditory cortex and receives input from it. In general, association areas are partly involved in such areas as learning, memory, thinking, and language.
3.1.2.2 Auditory where, what, and when pathways

Like the visual system, the auditory has been shown to have *where*, *what*, and *when* pathways (Schubotz et al., 2003). Clinical studies have shown that some patients with lesions have difficulties identifying where the sounds come from (Wortis & Pfeiffer, 1948; Sanchez-Longo and Forster, 1958; Klingon & Bontecou, 1966); but not until recent years research has confirmed that the lesion was on the *where* pathways. The auditory where pathway is in the dorsolateral premotor cortex; the *what* pathway is at the superior ventrolateral premotor cortex; and the *when* pathway is at the inferior ventrolateral premotor cortex. The location of pathways in the brain has shown on Fig. 4. In particular, research shows that when people determine the spatial position of sounds, the *where* pathway is activated; and the *what* pathway is activated when we identify sounds or determine which objects are sounding (e.g. telephone rings or doorbell rings). These pathways are generally activated in parallel. For example, when we hear the telephone ring, we usually also know where to pick up the phone and (usually) do not go to open the door. Another pathway is the *when* pathway. This pathway is essentially involved when we hear rhythmic patterns in music or analyze the temporal structure as a series of on-off tones (e.g. the Morse code).
The activation areas of these pathways are found in the frontal and the temporal lobes. They respond differently when one attends to different tasks. For example, when people need to localize an object, the where pathway is responsible for the task. However, unlike visual and somatosensory system, where spatial information is mapped directly from the sensory epithelia onto cortical areas, the auditory system contains a frequency map. Information about the location of sound is extracted/calculated from...
information that is sent from lower levels in the transmission, i.e. ears, brainstem, hindbrain, and midbrain.

Auditory pathways are very complex. Although they are part of five sensory systems of our body, we have much less knowledge about the auditory than about the visual system. There are many components in the sounds, such as pitch, timbre, rhythm, and other information that characterize specific sounds. However, as various research has shown, there seems to be no singular location in the brain for the identification of any of these components. It seems much more the communication between different activated regions of the brain that let us make sense of the sounds. That means, when we listen to a sound, a sentence, or a signal, all the different auditory pathways are working together and need to work together in order for us to perceive or ‘understand’ the sounds. Even small damage in any part of the auditory system could spoil one’s full understanding of sounds.
4 Elements of sounds and rhythm

4.1 Sounds: music and noise

Sound is a mechanical - frequently airborne - wave that can be characterized by its frequency, period, wavelength and amplitude. Energy level of the sound wave has to be strong enough to be aurally perceptible. However, physicists, audiologists, and musicians may define music and noise very differently. For example, in physics, a mechanical vibration may be considered noise if it is non-periodic and/or its frequency cannot be determined. For audiologists, any sound that has a large amplitude that cannot be tolerated by listeners may be considered noise.

Non-periodic sounds can be part of music in many cultures. For example, the mbira, a lamellophone from southern Africa, frequently has special noise generators (e.g. bottle caps) attached to it. The ‘noise-like sound’ is an integral part of many African music styles and their aesthetics. Some percussion instruments in Western music producing non-periodic frequency waves (e.g. cymbal, snare drums) are part of music making. Nattiez (1990) wrote in his book: “Just as music is whatever people choose to recognize as such, noise is whatever is recognized as disturbing, unpleasant, or both. The border between music and noise is always culturally defined--which implies that, even within a single society, this border does not always pass through the
same place; in short, there is rarely a consensus.” Therefore, there is no simple and universal concept of both noise and music. Non-periodic sounds can contribute to one of the most important features in music---rhythmic patterns. For example, the snare drum plays rhythmic patterns with its non-periodic sounds, and in vocal music the non-periodic sounds of consonants contribute to the formation and clarification of voice rhythms.

The term rhythm refers to the organization of events in time and it can be created by a variety of events like sound, light, touch, i.e. any temporally arranged sequence of events that stimulate the sensory organs. In my dissertation, I will only focus on rhythms that are produced by auditory stimuli. Before I describe musical rhythm, I would first like to describe the components of sounds that contribute to rhythmic formation.

4.2 The components of sounds

Sound is something we can hear and perceive through our auditory system; it is produced by pressure changes in one of three forms of matter---gas, liquid, or solid. Pressure changes in any of these media can be transformed into an electrical signal and presented in a visual form, which allows us to analyze the basic components of sounds, such as frequency, intensity, timbre, and duration. These components are described in the following.
4.2.1 Frequency and pitch

Frequency is the number of wave cycles or oscillations per time unit. The oscillations are measured in cycles per second, the unit of which is Hertz (Hz). Pitch, on the other hand, is our perception of the frequency of sounds. Normal human hearing covers the frequency range from ca. 20 Hz to 20000 Hz. However, different cultures describe pitch sensation in different terms. For example, in Western classic music, “higher pitch” indicates sounds with more oscillations whereas “lower pitch” indicates fewer oscillations. Other music cultures may have other terms, such as young and old (South American Indians) (Seeger, 1987), clear and muddy (Chinese)(Sun, 1980), etc. The examples show not all cultures express the sensation of pitch in similar terms; people express pitch sensation and pitch relationships with various, culture depended abstract terms or metaphors (see Zbikowski, 2002).

A succession of sounds with different pitches constructs a melody. Melodic contours are one of the most important components in most music. Vocal music is always based on some form of variation in melodic contours, with vocal music present in all cultures.

4.2.2 Intensity (Dynamics or amplitude)

The level of energy corresponds to the amplitude of the sound wave--- when the energy increases the amplitude gets bigger and vice versa. Intensity changes are perceived as changes in loudness and form the principle element of dynamics in music.
In music the dynamic changes can be sudden or gradual, sudden changes forming dynamic accents.

4.2.3 Timbre

Timbre relates to the waveform of a sound wave and to the quality or color of the sound. Waveforms are largely dependent on the instruments that produce the sounds. Every instrument has its own timbre. In addition, the timbre may vary with pitch or amplitude level. However, in Western music, musicians are trained to maintain the same timbre throughout the entire register of the voice or instrument although this is almost impossible to achieve. On the other hand, in many cultures (e.g. Chinese, Indian, Arabic, etc.), musicians recognize and accept the different timbres an instrument can produce, and even consider the different timbre of an instrument or the human voice to evoke in listeners’ different feelings and emotions during a performance.

4.2.4 Duration

Duration is simply the length a sound event from its onset to the end.

4.2.5 How the elements form rhythms

Rhythm refers to the arrangement of acoustic events in time, but Western music usually refers to it in the sense of “explicit divisions of time or space into intervallic time systems, recurrent and often characterized by periodicity, (Thaut, 2008).” However, the rhythm does not have to be explicit divisions. The idea of ‘explicit divisions’ of time
units originates from notation, not from actual music performance (Ong, 1982, Will, 1999).

Musical rhythm is the temporal organization of sounds. Rhythms are created by the time relationships between changing sound components. All sound features, pitch, duration, dynamics, and timbre contribute to the production of rhythm. For example, a rhythmic pattern may consist of tones with silent intervals in between. Another example is a single-pitched tone with dynamics changes. So, one or more elements that alter the sound could be enough to create a rhythmic pattern. How much modification of acoustic components is sufficient for our brain to detect the change, and what if the change is not sudden but gradual (e.g. a tone contour)? Gradual changes may not evoke a perceptible rhythm. It seems that rhythm is produced by more or less sudden changes that can be processed by our brain. On the other hand, if the rhythmic pattern is performed too fast, people may not perceive a distinct rhythm, because a sound event distance more than ca. 100 ms is required for rhythmic judgment.

Many studies (e.g., Weight, 2008, Scheirer, 1997) suggest that people can perceive rhythm automatically without working memory when the intervals between sounds are between 125 ms and 2000 ms, spanning what is called the ‘psychological present’ (Pöppel, 1978). In general, when the interval is smaller than 100 ms, people can hear the acoustic events but have difficulties processing the rhythm. If the interval is between 2 seconds to a few minutes, short term memory would be involved for
rhythm perception. If the interval were larger than a few minutes, the long-term memory would be involved.

Western educated musicians tend to realize or distinguish rhythm along the durational values that are available in notation. When musicians are asked to describe the rhythm of music, musicians usually do that in terms of notational concepts and values, because they have learned to consider notes as the units of rhythm. However, the embellishments of notes, vibratos, or even the noise from playing instrument (e.g. a saxophone’s finger cover noise and the ‘string gliding’ sound on a guitar) can also contribute to the rhythmic structure. Interestingly, people realize the noise when modern composers write these noises as part of the special techniques in the notation, but ignore the sound when it is not notated, even though the sound is always there.

4.2.6 Examples from world music

Not every music culture has notation, and most important, what we see on notation is not the same as what we hear, though it may correspond to what we expect to hear. “The idea that sound consists of independent, distinct events (for example, musical notes) is both useful and dangerous… musical meaning and even rhythm can also be conveyed by continuous shapes of time with no clear division into distinct events (London, 2006).” As an example from North Indian classical music, singing in melismatic style focuses on the melodic connections between pitches, so singers can improvise a melodic contour on one syllable. The pitches could be changed within one syllable, and some of the changes are very slow and gradual. Therefore it can be
difficult to analyze rhythmic structure, especially for outsiders. Take another example from my personal experience of erhu performance. A Western educated musician may ignore a small pitch change, such as a small glide or change in string pressure on one tone that can contribute to the rhythm patterns, because there is no such technique or ornamentation in Western music. They may hear it as a “wrong note” or even not realize the change at all.

However, this does not necessarily mean that Western people are worse at detecting small rhythmic variations than Indians or Chinese. It may simply be that, because they have never practiced these types of music, they are not used to detecting such changes or expect and pay attention to them, so that these types of changes mean nothing to them. Therefore, one’s sensibility of rhythmic patterns depends on one’s familiarity with the sounds and temporal structures of a specific music.

4.3 Beat, tactus, pulse, meter, and tempo

All music has rhythm, but beat, tactus, pulse, and meter may not always be present. They are precepts that emerge from our interactions with sounds. Even though beat, pulse, and meter, may not always be part of the physical structure of music, they are nevertheless important for rhythm perception, because our sense of beat or pulse is a by-product of entrainment and that can help us to perceive the rhythm (Jones, 1986). “Components of musical time divisions such as pulses, beats, and meter systems are relevant in this understanding of rhythm, (Thaut, 2008).” Pulse, beat, and meter are terms that reflect how people perceive or “feel” the flow of musical events.
The following section will briefly describe these terms.

4.3.1 Beat and tempo

The feel of beats is evoked by regular occurring acoustical events. This is also called pulse or tactus. More than two events are necessary for us to perceive a pulse. A series of regular pulses allow us to perceive the tempo. Tempo refers to the speed of pulse in the music. A regular pulse recurrence is necessary for measuring the tempo. However, the tempo does not indicate the ‘speed’ of music. For example, one of the sections in some Chinese opera has a rapid pulse but the main melody is sung slowly. The section is called ‘fast drum [beats] slow sing 緊打慢唱’. Therefore, a fast tempo needs fast repetition of pulses, but a fast tempo does not mean fast music. There are some musical genres where the pulse interval can be larger than 2 seconds, for example in the Indian music alap section. The tempo can still be regular and musicians are still counting with their hands (Clayton, 2008).

One should not confound pulse, tempo, and rhythm. Pulse and tempo are what we feel and rhythm is the temporal structure of the events what we hear.

4.3.2 Meter

Meter is an aspect of our engagement with the production and perception of acoustic events in time. Meter contains a number of regular beats or pulses and metrical units are formed by dynamical features. More than one cycle is necessary to evoke a
sense of meter. There are musics with beats but without meter, for example Indian Jor section, so meter is not necessarily found in every music performance.

4.4 **Rhythm beyond music, in particular speech rhythm and vocal music.**

The fact that humans are very good at detecting rhythm changes is obvious when we look at spoken language. Speech relies on small and rapid changes in acoustic features of consonants, vowels, intonations, tone contours, and accents. In most languages, the smallest speech unit can be very short, but are nevertheless still understandable and realizable. From these changes humans construct syllables, phrases, and sentences. When all the syllables, phrases, and sentences materialize in speech, this creates some kinds of linguistic rhythm. Various methods, e.g. specialized computer software (e.g. Praat), can be used to analyze linguistic rhythms.

Vocal music shares various features with speech and speech rhythm. However, unlike speech vocal music usually contains regular pulses, and the length of syllables and pitch may not be the same as in its speech forms. Unlike instrumental music, vocal music can make use of vowels and consonants, and produce various changes in acoustic features like intonation, timbre, tone contours, and accents. Vocalists can successfully imitate instrumental music, but it may be difficult if not impossible to use instruments to imitate vocal music. Numerous instruments in the world, e.g. sitar, erhu, have a sound spectrum close to that of the human voice, but they nevertheless sound like instrumental music. Once I tried to use computer software to convert vocal music to instrument timbres (violin, flute, oboe, erhu). The output was an instrumental sound,
that nevertheless still sounded like vocal music, because features of the human voice were still present, and it was those features that made the converted file sound like vocal music. Thus, the difference between vocal and instrumental music (rhythm) is not only the (static) timbre, but there are several elements that distinguish the two. Especially the spectral dynamics of the voice allow us to tell vocal from instrumental sounds.

To conclude this chapter: any change of sound components can contribute to the rhythm structure. However, humans may process rhythms differently depending on the time intervals between acoustic events. Long term or short term memory may be applied when processing rhythms, but if the interval is too small (less than circa 100ms), people may not be able to process the rhythm.

The following chapters present research on musical rhythms that only require short term memory. The results from a cognitive study with a set of experiments will be reported and discussed in order to understand whether there are differences in processing of vocal and instrumental rhythms.
5 Rhythm perception experiments

5.1 Introduction—Previous studies on music and language

The human auditory system allows people to receive and hear sounds from about 20 Hz to 10000 Hz. Within this range, people are able to perceive human voices and other sounds, such as environmental sounds or the sounds of other animals. The question of whether people listen to different kinds of sound in the same way or not has been discussed by many research teams (Bent et al., 2006, Levy et al., 2001, Vouloumanos, 2001, & Zatorre et al., 2002).

The currently dominating view in neuroscience concerning the relationship between language and music processing by the brain is that they are processed by different neuronal networks (Janata, 2002; Levitin & Menod, 2003; Koelsch, 2002; Tillman, 2003; Hebert et al. 2003). Some studies show that the left dominance hemisphere is for linguistic patterns whereas the right hemisphere is for music processing. Several studies point to a domain specificity of both speech and music processing (e.g. Zatorre et al., 2001.) On the other hand, some studies advance evidence that music and linguistic information are processed in interaction and by similar and overlapping brain areas despite the fact that speech and music utilize different spectral dynamics (e.g. Schoen, 2005).
In the study by Schoen et al (2005), their first experiment compares the linguistic and music information in songs. The idea of using songs for research is that singing makes a nice connection between language and music. They had their research participants attend to different acoustic parameters (language or music) and then observe their neural activation pattern during an fMRI registration. They found that language and music processing involve similar cortical regions, and lexical information and pitch contours processing cannot be processed separately. Their study shows how important language and music processing relate to each other. On the other hand, Belin (2000) shows that the vocal sounds lead to a greater neuronal activity in the superior temporal sulcus. Belin’s research does not contradict Schoen et al’s results. But Belin shows that the human voice contains unique information about speakers’ identity and emotional state.

5.2 Previous studies on rhythm processing

Most dance forms around the world are closely coordinated with the rhythm or beats of their accompanied music. This phenomenon happens not just by accident, but is most likely based on some biological-physiological principles. Studies have shown that several motor areas respond when the brain processes rhythmic information, even without any body movements, the motor areas in the brain can still be activated. As has been shown by Grahn and Brett (2007), some motor areas are still activated when rhythmic patterns are perceived. In particular, their research has shown that basal ganglia and supplementary motor area are activated when regular beats are present in the stimuli.
Although Basal ganglia and supplementary motor areas also have other functions, they are critical for temporal sequencing (Shima and Tanji, 2000). These two areas may be involved in isochronous beat interval detecting (Grahn and Brett, 2007). fMRI data also shows bilateral activities in the above areas are increased when listening to simple rhythm. This implicates that a bilateral network in the Basal ganglia and supplementary motor areas is activated in rhythmic processing (Grahn and Brett, 2007).

The ability to reproduce rhythms improves significantly when rhythm occurs with regular beats (Patel et al., 2005; Essens and Povel, 1985). In addition, the reproduction is more accurate for metrical than for nonmetrical rhythms (Essens, 1995; Sakai et al., 1999). It is claimed by many scholars that integer ratio rhythm and simple rhythm are straightforward and easy to be perceived, produced, and reproduced. This can be demonstrated with many types of Western music, especially popular music. Examples can be found in popular music, folk songs, and lullabies. However, Western music education tends to train people to produce rhythms based on small integer ratios. Because most studies on rhythm were done by Western scholars and participants, it would be relevant to know more about perception, production, and reproduction of rhythms by people in the rest of the world.

Psychological studies usually propose two types of rhythm: small integer ratio or non-integer ratio. However, people tend to regularize non-integer ratio rhythm into integer ratios (Grahn and Brett, 2007). This suggests that musicians tend to perform small integer ratio rhythm better than non-integer ratio rhythm. And it can also be
assumed that music performance tend to have more small integer ratio rhythm than non-integer ratio rhythm. Yet, a study on Australian aboriginal music showed that musicians in real performances do not show a preference for simple integer ratio rhythms (Will, 2005). This result challenged many previous studies that integer ratio rhythm is more dominant in human music societies, and a perfect integer ratio may not always show in real performances. Perhaps there is no “internal clock” in our mind, brain, or body, one like a metronome that serves a reference of perfect integer ratio rhythm to us.

Many researchers also investigated another type of rhythm processing: small integer vs. non-integer ratio rhythm. Previous research has shown that the prefrontal and parietal cortices are the main areas for temporal perception and decision-making. Clinical studies also showed that patients with lesions in these areas show rhythmic deficits (Brust, 1980; Mavlov, 1980; Polk and Kertesz, 1993). On the other hand, Sakai et al. (1999) suggest that there are two patterns of brain activations. In short, small integer rhythm, which they called metrical, leads to more activation in the left premotor, parietal areas and right cerebellar anterior lobe; whereas non-integer rhythm, which they called non-metrical, seems to lead to more activation in the right prefrontal, premotor, parietal areas, and bilateral cerebellar posterior lobe. Thus, they suggest that there are two channels for rhythmic processing, one for integer and one for non-integer rhythms. However, the two activation patterns were not statistically compared, as Grahn et al. (2007) already pointed out in their article. Yet, it is unknown how much difference exists between these two rhythms; the difference they found may not be significant because they did not test it. It could be that the left hemisphere did activate for both
conditions but the right hemisphere showed stronger activation during the nonmetric rhythm; and it is also possible that the activation on the right hemisphere is not due to rhythm itself only, but other factors. For example, the activation could be due to the unfamiliarity of rhythm. Given that prefrontal cortex, posterior lobe, and right frontoparietal areas are involved in working memory and attention, it suggests that nonmetrical rhythm may “require additional monitoring processes (Sakai et al, 1999).” As other research has shown, people in Western cultures seem to be more familiar with metrical rhythm and it is not surprising that “additional monitoring processes” are present when they listen to nonmetrical rhythm. Nevertheless, left premotor and parietal may always activate when listening to rhythm in the Sakai et al. study.

5.3 Introduction to the present research question and its experiments

Various psychological and physiological studies show that the human voice occupies a special place in terms of cognitive processing. For example, in a series of word repetition experiments we have shown that vocal melodic contours evoke larger facilitation than instrumental contours. In the current study I investigate whether there is also a difference in cognitive processing of instrumental and vocal rhythms. The study will be performed with native English speaking musicians. One of the central questions that is addressed here is: do humans listen and respond to vocal rhythm and instrumental rhythm in the same way? In the first experiment the question will be investigated through reaction time measurements in rhythmic decision tasks. The outcome of this study will be relevant to the current debate on rhythm perception and
processing by the brain, in particular, for the question of whether there is only “one type” or “several types of rhythms” in human cognition. The outcome may also contribute to a better understanding of the evolution of music.

5.3.1 Materials

Ten short excerpts of 1.5 to 2 sec length (mean 1.73 seconds), sung by male performers, were selected from Dyirbal field recordings of R.M.W. Dixon (published on Larikin LRF 378 CD “Yirbal Song poetry”.). From the CD we also sampled one clapstick sound as a ‘standard’ with which we either replaced the originally recorded clapstick sounds or created new accompaniment for the excerpts that were solo voice recordings. The interval ratios of the newly created clapstick pattern followed a distribution established through measuring the sequential ratios of all clapstick intervals on the above CD. From each of the 10 excerpts we created three variant versions through editing with CoolEdit: one version with a changed voice rhythm, one with the clapstick rhythm changed, and one version in which both the clapstick and voice rhythm were modified. Rhythmic variants were created by changing the timing, inserting or deleting one or two sound events in either the voice or the clapstick track or both (for the third variant). Variants were always only one change in voice, but the clapstick may have one or two changes. This resulted in 40 variants including the original excerpts. The sequential ratios of the vocal and clapstick rhythms in these variants are shown in fig. 5. According to the local tradition there are more events (6.7 per variant) and hence more ratios in the voice than in the clapstick patterns (4.3 events/variant)
Figure 5: Distribution of sequential time interval ratios for clapstick (Cl, top) and vocal (Vo, bottom) rhythms in the 10 musical excerpts. X-axis: interval ratios; Y-axis: number of occurrences.

With these variants we then formed 4 stimulus pairs for each excerpt by combining two versions: one in which voice and instrument rhythms were identical (S_{voice}S_{instr}, e.g. one version and its repeat), one in which voice rhythm differed and instrumental rhythms were identical (D_{voice}S_{instr}), one in which instrumental rhythms differed and the voice rhythms were identical (S_{voice}D_{instr}) and one in which both rhythms were different (D_{voice}D_{instr}). The two versions were aligned in such a way that in half of the stimulus files the clapsticks of the second excerpt started ‘on beat’ with the clapstick pattern of the first excerpt, and in the other half the second excerpt started off beat (around 2/3 of the beat/pulse of the first excerpt). This resulted in a set of 40
stimuli of sound pairs, 10 for each of the four stimulus conditions ($S_{\text{voice}}S_{\text{instr}}$, $D_{\text{voice}}S_{\text{instr}}$, $S_{\text{voice}}D_{\text{instr}}$, and $D_{\text{voice}}D_{\text{instr}}$) with half of the pairs aligned ‘on beat’ and half ‘off beat’.

The reasons for choosing these recordings for our experimental stimuli are the following.

First, as we only want to investigate rhythm processing, we want to avoid any lexical or other linguistic effects of the stimuli on rhythm processing. Linguistic research has shown that participants respond to a language differently based on the familiarity with the language. In previous studies we eliminated the segment information from the language in order to avoid the lexical effects (Will and Poss, 2008). However, this type of modification may severely alter the rhythms of the stimuli and their intelligibility. As the differences between vocal and instrumental rhythms are the central element in our experiment, we need to keep the rhythms as natural as possible and the vocal rhythms should not be modified. However, as we want to exclude lexical effects, we chose a language that none of the participants in our experiment is familiar with and that is phonetically distant to avoid interference of unwanted associations. We therefore chose an Aboriginal Australian language as stimulus material.

Second, unlike most of the singing in Western classical music, the music of these recordings is not based on instrumental music but vocal music. Western music, even if it is vocal music, is strongly based on instrumental melodic and rhythmic structures. As our research question centers on the contrast between vocal vs.
instrumental music, we wanted to select a singing style that shows no or almost no influence from instrumental music.

Moreover, for this particular singing style it is cultural practice that it is accompanied by clapsticks, which allow us to research the relationship between vocal and instrumental music from the same recording. So the experiments would be based on stimuli generated from the real music examples, not sounds artificially created in the lab. Consequently, the ‘instrumental sounds’ in our experiment are limited to clapsticks only. Surely one should not ignore other instruments, both melodic instruments and percussive instruments, and future research should continue exploring the rhythm processing of other instrumental sounds in order to show whether the results we obtain here only hold for clapstick rhythms or for instrumental rhythms in general.

In addition, we adjust the ISI (inter-stimulus-interval) based on the on-going pulse (beat) of stimulus 1, in order to control whether the on-going pulse effects participants’ decisions with the task. There are two conditions: on-beat and off-beat conditions. In the on-beat condition, the duration between the last beat of stimulus 1 and first beat of stimulus 2 is the same or twice as long as the duration of beats in stimulus 1. So that in the on-beat condition, participants would feel the pulse from stimulus 1 continuing in stimulus 2. On the other hand, in the off-beat condition (fig. 7), the duration between the last beat of stimulus 1 and first beat of stimulus 2 is different from the duration of beats in stimulus 1. As every sound stimulus has two rhythms: vocal and instrumental rhythm, the measurement for the adjustment of the inter-stimulus interval
(ISI) is the instrumental rhythm. We use the instrumental rhythm as the base line because the instrumental rhythm has more regular rhythm than vocal rhythm. The beat alignment of the two sound files in each stimulus will test whether the alignment had any effect on the responses.

Figure 6: One trial presentation.
Figure 7: Waveform of one stimulus trial. X-axis: time in seconds; Y-axis: relative amplitude (in dB). The two stimuli are aligned in the off-beat condition; clapstick rhythm are different, vocal rhythm are the same. The beat of Stimulus 1 is .63 second, the duration between the last beat of stimulus 1 and first beat of stimulus 2 is 1.09 seconds.

5.3.1.1 Real world stimuli

The present study is one of few studies to use real world stimuli. In most of psychological studies, stimuli are usually made in the lab and not taken from real performances. As there is no music performance with an absolutely perfect integer ratio rhythm (except some computerized music), the research with these ‘perfect’ stimuli may be able to demonstrate how we listen to these ‘stimuli’ but not necessarily to real ‘music.’ It is always a debate between psychologists and musicians how much these stimuli represent actual music. Though these laboratory stimuli can hardly be called
music, there is no doubt about the contributions to music cognition from these studies; they give us important and fundamental knowledge on how humans process various types of rhythm and sound.

In this present study, although some of the stimuli were manipulated in the lab, all our original stimuli were from live recordings. All manipulated sound files have only one point in time that was changed from the original, and we try to let our stimuli be as close to the real performance as possible. These types of stimuli allow us to get closer to the study of how human process actual performed music.

One of the reasons that other studies had to use artificial stimuli was that researchers wanted to have as much control over the stimuli as possible throughout the experiment. In rhythm studies, usually the only element that changes throughout the experiment is the interval between stimuli; other elements such as pitch, dynamic, timbre, and intensity would be consistent and identical among the stimuli. This is to control as much as possible for any confounds.

However, we used two completely different sounds in our experiments. Our two types of sounds, vocal and instrumental sounds, have very different sound features. In fact, unlike other studies that try to avoid changing more than one element, we took the difference as an advantage in our experiment. Because the sounds were so different, we could play both stimuli at the same time and each stimulus was still recognizable by participants. We played each sound presentation (i.e. vocal and instrumental sounds) twice for two different tasks, so participants heard both stimuli in both tasks.
This design could provide a suitable methodology for related studies. For further research it would be interesting to repeat these experiments with other instruments, especially with melodic instruments.

5.3.2 Participants

Only English-speaking musicians participated in this experiment. They were recruited from the student population at Ohio State University. All reported normal hearing, speech, and language. To avoid any language effects, any effects could expect, all of them speak English as their native language. As we do not know if music training would affect the results, all participants are musicians. In addition, it is easier to explain the task to musicians than non-musicians, in particular, when applying the term ‘rhythm’, ‘tempo’, or ‘timbre’. As this study is not only performed by behavioral experiment but also fMRI, we want to make sure all participants understand the instruction; because it is very costly if participants cannot perform the task and have to repeat the experiments. All were graduate or undergraduate music major students at the School of Music at Ohio State University. Their area of concentration included performance, composition, and musicology. All musicians play at least one instrument and still continue to practice regularly.

5.3.3 Experiment procedure

Interested and eligible participants came to the research location; they were introduced to the research facility and procedure of the experiment. Following these explanations they decided whether to participate or not. If they decided to participate
they were asked to sign the consent form. The participants clearly understood that they could terminate the procedure and exit the study at any time. After they signed the consent form, they received a tutorial section of the experiment.

In the tutorial section, the task, procedure was explained to them, and they were familiarized with the research environment (especially for fMRI). They practiced on some trials that were not part of the experimental stimuli, and we would make sure they obtained correct answers, or at least most correct, before they started the experiments.
6 Behavioral experiments and results

6.1 Reaction time (RT) measurements

Reaction time (RT) measurements are quite common in psychological behavioral research. Reaction time is the time interval between the presentation of stimuli and the subsequent behavioral response. Stimuli can be any sensory stimulus, for example, auditory, visual, or touch, etc. The measurements can indicate the speed of mental processing, i.e., how much time the participants need to perform the mental operations for the experimental tasks, and factors that influence this processing. The responses are usually recorded as button presses, vocal recording, eye movement, or other behavioral measures.

There are two time points in the reaction time measurements: the point when participants are presented with the stimuli and the onset of the behavioral responses. So, the question of where to begin and end can be very demanding, especially the measurements may affect the results if the starting or ending points are altered. The ending points of measurements, or response time point, are always the point that participants make behavioral response such as button press, eye movement, or vocal response.
Depending on the nature of research, when to start measuring the onset of reaction time is not always predefined. If the stimulus is presented at once, as is the case e.g., with visual images, the onset of measurement is typically the same as the onset of presentation. Another type of stimulus like a non-changing acoustic stimulus presentation, the onset of measurement usually starts simultaneously as presentation. For example, in an experiment of absolute pitch identification, where none of the sound elements (timbre, pitch, amplitude) is changed throughout the whole presentation, the onset of measurement can be the beginning of the presentation.

On the other hand, if the acoustic stimuli are changing in time, and if the decision task is related to these changes, it can be a challenge for researchers to decide where to start the onset of reaction time. In the present study we defined reaction time (RT) as the time between the point of difference and the time of button press. Points of difference are those time points where the second stimulus differs from the first. For the same-condition this point is the onset of the final event in the rhythmic pattern, i.e., either the last clapstick sound (for instrumental rhythms) or the last syllable (for vocal rhythms).

Trials with errors will be considered in the error analysis but are excluded from the reaction time analysis. Errors include (1) wrong decision, (2) press button earlier than the points of difference, (3) press button later than the end of decision period.
6.1.1 Reaction time and participants’ response strategies

Participants were instructed to indicate their decision on whether the vocal or clapstick rhythms were the same or not as quick and as accurately as possible by pressing the appropriate response button and, if they detected a difference not to wait until the end of the sound file. However, even if the instruction is to press button as soon as possible, we have no control if all participants follow the instruction. It is possible that some participants had other strategies; perhaps they may wait until the end of whole presentation and then press the button. Thus, the RT from those participants would be irrelevant for our research question. Therefore, the following analysis was performed to control whether our participants follow the instruction.

If they follow the instruction, their reaction time should somehow correlate with the point of difference, because the instruction is to press the button as soon as they hear...
the difference, or the end of sound file if the two sound files are the same. For example, if the point of difference appears earlier in the files, their responses should also be earlier. If, however, their responses always occur after the second sound, i.e. regardless of the time of point of difference, then their responses are not according to the instruction. As the point of difference may appear anywhere (i.e. early, middle, or late) in the second sound file, we expect participants’ responses to vary accordingly.

We tested whether our participants understood and followed our instruction correctly in the following way. We grouped our point of difference in three classes according to when the difference point occurs. If the point of difference appears between 0 and 0.7 second, they are in the early class (A); if the point of difference appears between 0.7 and 1.2 seconds, they are in the middle class (B); if the point of difference appears after 1.2 seconds, they are in the late class (C) (Fig. 9). We expect their response times to correlate with these three classes: early, middle, and late.

Therefore, we performed an ANOVA analysis based on these three groups in the relation to participants’ RT. The result shows highly significant correlation (p=<.0001) and indicates that, overall, the participants did follow our instruction when responding.
Figure 9: ANOVA of the reaction time with difference-point classes as factors. Factor levels: early, middle, and late difference-point. This plot graph shows that early class (A) elicits earliest response, followed by middle class (B), and then late class (C) difference-point. X-axis: difference-point classes. Y axis: response time from the start of the first sound (in milliseconds).

Note that A and B contain ‘different’ answers, and C includes all the ‘same’ answers and a few ‘different’ answers. In addition, the standard deviation of C is clearly smaller than A and B. This may be an indication that participants process the two different conditions (same and different) differently. The further analysis will be discussed in the following sections.
Figure 10: The distribution of points of difference in the stimuli. The three subdivisions for the test were: A, from 0-0.7 (seconds) with the peak around 0.3. B, from 0.7-1.2 with a peak around 0.9. C, from 1.2-2 with a peak around 1.5. Again, C has more counts than A and B because C includes all ‘same’ answers and a few ‘different’ answers. X-axis: time (in seconds) and Y-axis: the number of occurrences.

6.1.2 Stimulus classification into small-integer and complex-integer ratio pattern

We classified the rhythmic patterns as small integer ratio or complex ratio patterns as follows: the ratios for all successive rhythmic intervals were calculated. If the respective durations fell within ± 5% of the durations for ratios of 1, 0.5, 0.33, 0.66, 0.75, or 0.25 (corresponding to interval proportions of 1:1, 1:2, 1:3, 2:3, 1:4, and 3:4), they were classified as small integer ratios, the remaining ratios were classified as complex. The rhythmic patterns were then labeled according to whether they had more small integer ratios (s) or more complex ratios (c). There were 12 s and 28 c pattern for
the vocal rhythms and 23 s and 16 c patterns for the clap rhythms (one clap pattern could not be classified this way because it contained only two events).

That means, for example, the rhythmic intervals have been classified as ratios of 1:4 may comprise a range of ratio from 0.23 to 0.29. Other small integer ratios range as following: 1:3 (from .31 to .37), 1:2 (from .47 to .53), 2:3 (from .63 to .69), 3:4 (from .73 to .79), and 1:1 (from .97 to 1.10). Any ratios outside of those ranges are complex integer ratios.

For the 40 stimuli, each then is classified into either ‘small integer ratio group’ or ‘complex integer ratio group’ based on the frequency of appearance of small integer ratio and complex integer ratio rhythm. For example, a stimulus has more small integer ratios than complex integer ratios, then this stimulus is small integer ratio group.

Example from one of the stimuli,

Time points of each rhythmic event in timeline (unit: second)

| 2.9 | 3.11 | 3.32 | 3.58 | 4.1 | 4.24 | 4.45 |

Intervals between two events (e.g. 3.11-2.9=0.21).

| 0.21 | 0.21 | 0.26 | 0.52 | 0.14 | 0.21 |

Ratios for each interval (e.g. 0.21/0.21=1).

| 1 | 0.808 | 0.5 | 3.714 | 0.667 |
However, if the ratio is larger than 1, it would be divided by 1. (e.g., 1/3.714 = 0.269)

\[
\begin{array}{cccc}
1 & 0.808 & 0.5 & 0.269 & 0.667
\end{array}
\]

Finally, four out of five ratios are small integers, so this stimulus is in the group of ‘small-integer ratio’.

### 6.2 Participants

Fourteen English-speaking musicians participated in this experiment. Participants included seven females and seven males (mean age = 26.2 years) musicians.

### 6.3 Task procedure and design

The 40 stimulus pairs were arranged in two presentation blocks of 20 stimuli and each block was presented twice, once with the instruction to decide whether the clapstick rhythm of the two excerpts was the same and once to decide whether the vocal rhythm were the same. Stimulus presentation was controlled by DMDX (Forster & Forster, 2003) and participants listened to the stimuli sequences via headphones. Each participant was given a brief practice session with 5 stimulus pairs that were not part of the experimental stimuli. Before the start of a block the respective task was displayed on the computer screen and remained visible until the end of each block.

The screen would show either “is vocal rhythm the same?” or “is clap rhythm the same?” After they understood the task, they press the space bar on the computer and started the task. In one task participants were instructed to ignore the vocal rhythm and decided whether the clapstick sounds are the same or not when they see “is clap rhythm
the same?” By following the second task of the experiment, the participants were
instructed to ignore the clapstick rhythm and decide whether the vocal are the same or
different when they see “is vocal rhythm the same?”

They then listened to pairs of sounds and decided if the rhythm in both sounds
was the same or not. They were instructed to press the “Yes” button immediately after
they heard any difference or wait until the end of sound presentation to press “No”
button if they did not hear any difference. They should press the button if they detected
any differences when the sound presentation was still playing, and if they did not hear
any difference, they press the button right after the sound presentation finished or
during the rest period. Participants’ responses of decision making were recorded via
computer mouse pad. Right click represents “Yes” and left click represents “No.” The
“Y” and “N” were labeled on the mouse pad. They were told to use right index for “Y”
and right middle finger for “N.” Their reaction time was recorded by DMDX. After
button pressing, they would need to wait about 2 seconds before the next sound
presentation play and then they would repeat the same procedure again for the next
presentation (Fig. 11). The order of presentation of the four blocks was balance across
subjects.
6.4 Analysis

6.4.1 Error analysis

Due to an initial miscomprehension of the instruction that was corrected after the second run, one subject showed an error rate of over 50% and was excluded from the error analysis.

For the remaining 13 subjects the mean correct response rates was 79%, ranging from 74 to 88%. The overall error rate is not independent from the stimulus conditions (Chi square = 19.90, p= 0.0002); there are significantly fewer errors in the ss condition (12%) than in the other three conditions (dd=24%, ds=26%, sd=21%). Table 1 shows the actual error frequencies in different conditions.
Table 1: Error frequencies for the four conditions. Y=error, N=correct. ss = both rhythms same; sd = voice same, clapstick different; ds = voice different, clapstick same; dd = both rhythms different.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Y</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>dd</td>
<td>197</td>
<td>63</td>
<td>260</td>
</tr>
<tr>
<td>ds</td>
<td>192</td>
<td>68</td>
<td>260</td>
</tr>
<tr>
<td>sd</td>
<td>205</td>
<td>55</td>
<td>260</td>
</tr>
<tr>
<td>ss</td>
<td>230</td>
<td>30</td>
<td>260</td>
</tr>
<tr>
<td>Totals</td>
<td>824</td>
<td>216</td>
<td>1040</td>
</tr>
</tbody>
</table>

Testing the responses for clapstick and vocal rhythm decisions separately, the error distribution for the clapstick rhythms shows no influence of the stimulus condition (chi square = 3.15, p = 0.37). However, as shown in the Fig. 12, there is a significant dependency between error distribution and stimulus conditions for the vocal rhythms (chi square = 12.32, p = 0.0064), with 8.5% errors in ss condition and 25% in the dd condition. The error rate for the vocal rhythm decision increases from the ss to the sd condition, where vocal rhythms are the same but the (task-irrelevant) clapstick rhythms are different, to the ds condition where vocal rhythms are different but clapstick rhythms are the same, and reaches a maximum in the dd condition, where both rhythms are different.
Figure 12: Mean % correct responses across subjects for stimulus conditions and clapstick (c) and voice (v) task. Error bars = 1se. ss = both rhythms same; sd = voice same, clapstick different; ds = voice different, clapstick same; dd = both rhythms different. X-axis: conditions; Y-axis: mean correct responses (in %).

As the vocal stimuli contained more small-integer than complex clapstick rhythms (s. Material & Methods), the different error rates for the two rhythm types could be due to these different distributions of complex and small-integer ratio rhythms. However, a chi-square test for errors and ratios did not reveal any dependencies of error rates on rhythm ratios (chi-square = 0.402, p = 0.53).

6.4.2 Reaction Time (RT)

Analysis of RT by repeated measure ANOVA shows significant effects of stimulus conditions (F(3,13)=108.19; p<0.0001) and the interaction between stimulus condition and task (clapstick/vocal rhythm) (F(3,13)=97.54; p<0.0001).
However, as shown in Fig. 13, in contrast to the results on error rate, the non-task relevant rhythm has no significant effect on RT and we can simplify the model by only considering the same / different condition of the task-relevant rhythms. Again we find a significant effect for stimulus condition (F (1,13)=321.93; p<0.001) and for the interaction between stimulus condition and task (F (1,13)=6.79; p=0.016)(see fig.14). Post-hoc (Bonferroni/Dunn) analysis shows there is no significant RT difference for clapstick and vocal rhythms when the rhythms to be compared are different (mean RT for vocal: 1210 ms; for clapstick: 1200 ms). The interaction arises from the different reaction times in the same condition: here RTs for vocal rhythms are significantly faster than for clapstick rhythms (mean RT for vocal: 308 ms; for clapstick: 503 ms; mean difference 195 ms, p<0.001).

Figure 13: Interaction graph for stimulus conditions and task condition. Error bars = 1 standard error (se.). c = clapstick rhythm task; v = voice rhythm task. ss = both rhythms same; sd = voice same, clapstick different; ds = voice different, clapstick same; dd = both rhythms different.
Figure 14: Interaction graph for task condition and task-relevant stimulus condition. Error bars = 1se. s = task rhythm same; d = task rhythm different; c = clapstick task; v = voice task.

As clapstick and vocal rhythms are independent patterns there is a possibility that this result may be due to different end points of the two rhythms. If, for example the last clapstick events were consistently earlier than the last vocal event (syllable), the above result could find an explanation if, in the same-condition, subjects waited until the end of the sound file to give their responses. As RT is measured from the point of difference, here the final event, the RT difference for vocal and clapstick rhythms should correspond to the mean difference between the final events in both rhythms. However, in designing the clapstick rhythms we took care of a balanced distribution of the last events in the same condition. Fig.15 shows the differences between the end of sound (eos) and the last event for the clapstick and vocal rhythms in the same-condition. The mean for the 20 vocal rhythms is 237 ms and for the 20 clapstick rhythms is 238 ms.
Clearly, differences for the time of occurrence of the final events in the same condition cannot explain the highly significant difference of 195 ms between the vocal and clapstick rhythms decision in the same condition.

**Figure 15**: Distribution of point-of-difference for vocal (v, red dots) and clapstick (c, blue dots) rhythms in the same condition. Y-axis: time difference between the end of the sound file (eos) and the onset of the final event in the rhythmic pattern.

### 6.4.3 The on-beat/off-beat effect

In this experiment we had also included an on/off beat alignment of the two sound files in each stimulus in order to test whether the alignment had any effect on the responses. We did find a significant effects of the alignment ($F(1,12)= 11.923$, $p=0.0048$), with the response times in the on-beat condition being shorter (753 ms) than in the off-beat condition (853 ms) (fig.16)
However, closer inspection of the stimuli features showed that although most features like location of difference points or the number of clapstick/vocal decisions were balanced across on/off conditions, this was not the case for the number of same/different decisions. There were more ‘same’ stimuli in the on than the off condition (121/97) and vice versa for the ‘different’ stimuli (90/144). With the same condition having a considerable advantage over different condition, this may confound the alignment effect. To test whether the alignment effect survived when this confound was eliminated we performed the following experiment.

6.4.3.1 Experiment 2

The identification of an alignment effect for rhythm that only lasts two to four beats would be a novel and potentially significant phenomenon that merits further study.
We therefore performed a second experiment in which we controlled for potential confounds by presenting all stimuli in both on- and off-beat alignment.

### 6.4.3.2 Participants

Eight music major students (four female), none of whom had participated in the first experiment, took part in this study.

### 6.4.3.3 Material

From the 40 stimulus files of experiment 1 we randomly selected 20 and created two sets for the vocal rhythm task by aligning the two sound excerpts of each pair ‘on beat’ for one set and ‘off beat’ for the second set. For the clapstick rhythm task we took the remaining stimulus files and created another two sets in the same manner.

### 6.4.3.4 Procedure

Again participants were given instructions to decide as quickly and as accurately as possible whether the clapstick or vocal rhythms of the two excerpts in each stimulus were the same, and subjects had to press the ‘yes’ or ‘no’ pad button to indicate their decisions. The order of presentation of the four blocks was balanced across subjects.

### 6.4.3.5 Results

For the reaction time we obtain $F(1,7)= 1.67$, $p = 0.24$, i.e. the on/off beat condition does not significantly affect the reaction time (Fig. 17). Results from the previous experiment were obviously due to the imbalanced distribution of the same/different stimuli in the two conditions.
This result is corroborated by the error analysis. The error rate is independent of the on/off alignment (Chi square 2.04, p = 0.15) (Fig. 18).
Although the two decision tasks in this experiment were performed on different subsets of stimuli we basically obtain similar results for the stimulus conditions as in the first experiment. The error rate for clapstick decisions does not vary significantly across the stimulus conditions, while there is again an increase in errors across the conditions for the vocal decisions with a significant difference between the ss and dd condition.
Figure 19: Mean correct responses (in %) across subjects for stimulus conditions, clapstick (c, blue), and voice (v, red) task in experiment 2. Error bars = 1 standard error. ss = both rhythms same; sd = voice same, clapstick different; ds = voice different, clapstick same; dd = both rhythms different.

6.5 Discussion of errors

In the error rate analysis we found the error rates were different for the different task conditions. Overall error rate was 21% and is dependent on the stimulus conditions. Error rate for ss was 12%, which was lower than dd (24%), sd (21%), and ds (26%).

The clapstick rhythm error rates did not seem to be influenced by stimulus conditions, because error rates were more or less stable across conditions (the existing differences were not significant). Vocal rhythm error rates, on the other hand, were significantly influenced by the stimulus conditions.
It seems the level of difficulty for clapstick rhythm was similar in both the same (19%) and different (15%) rhythm tasks; participants did not have more difficulty on identifying one condition than the other. But participants seem to have much more difficulties with the identification of vocal rhythm in the different task (32%) than the same task (15%).

However, there could be other reasons for the different error rates. For example, the degree of complexity or regularity of the rhythm. This can e.g. be defined by its ratio: small integer ratios constitute regular rhythms whereas non-integer ratios makeup the irregular rhythms. Previous research (Sakai et al., 1999) has suggested that the perception of integer ratio rhythms may be different from that of non-integer ratio rhythms. In particular, the authors propose that there are different neural representations for each of these rhythms. In addition, Essens (1985) showed that people seem to have a preference for the integer ratio rhythm, and participants tend to transform the non-integer ratio to the integer ratio rhythm in reproduction experiments.

Therefore we test whether there was a correlation between the errors and ratio groups by applying chi-square test. But no significant correlation between the two factors was found (p= 0.53). So rhythm ratios did not appear to influence participants’ error rate in our experiment.

Another point to address here is the significant difference between errors for clapstick and vocal rhythms in the dd condition and the fact that clapstick rhythms have the lowest error rate in this condition. Our hypothesis is that this might be due to a
design imbalance between the two rhythms; while vocal rhythms in this condition differ only in the timing of one event, 70% of the clapstick rhythms (7 out of 10) differ in more than one event. This could lead to the different clapstick rhythms being more easily detected (i.e., larger difference to the reference rhythms) and thus showing a reduced error rate. This hypothesis will be tested in a future experiment.

6.6 Discussion of the reaction time measures

The results of this experiment provided an interesting contrast between the two types of rhythm processing. Unlike in the error analysis, the reaction time did not seem to be influenced by the task-irrelevant rhythms. Therefore, we could simplify and only look at the two main conditions: same and different task rhythms. In this experiment, we found reaction times were faster for vocal than for clapstick rhythm decisions in the same condition, but their reaction times were not significantly different in the different condition. In this condition the variability and error rates for clapstick rhythm decisions were smaller than for vocal rhythm, which seems to suggest that the clapstick rhythms were easier to identify and to process. However, before exploring these results we need to address the considerable reaction time differences for the same and different condition.

Differences between same-different judgments have been found in numerous experiments (e.g., Proctor, 1981, for overview) and a number of theoretical proposals have been made to explain the observed reaction time disparity. In Cohen & Nosofsky’s (2000) exemplar-based random walk model (EBRW), two component processes are
assumed to drive *same-different* judgments. First, when a pair of objects is presented, observers are assumed to make on-line judgments of perceptual similarity. High values of judged similarity tend to drive a random-walk counter toward a *same* response criterion, whereas low values of judged similarity tend to drive the random walk toward a *different* response criterion. One key prediction of this model is that it should be more difficult to respond *same* to an identical object pair that resides in a dense region of the multidimensional similarity space than to a pair from an isolated region. However, the prediction of this model, that was constructed to explain results from matching of simultaneously presented visual stimuli, is contrary to our results. In addition, the way we had constructed our stimuli did not create regions of different density: each stimulus had four variants, and each of the variants differed from the others by only the timing of one or two single events. It therefore seems unlikely that the reaction time differences between same and different judgments can be explained by processes underlying the EBRW model.

Proctor (1981) proposed a unified theory for matching, priming and repetitions tasks. The difference in reaction time to *name-same* and *physical same* pairs has been attributed to differences in the level of processing at which the match is made (Posner, 1978). Based on his experiments with visual stimuli, Proctor concludes that all sequential matches are apparently based on name codes, with facilitation affecting the rate at which the name of the second stimulus is obtained. The unified model he proposed is one in which *different* reaction times for both simultaneous and sequential stimuli vary as a function of the similarity of the *different* stimulus pairs. The primary
difference between sequential and simultaneous matches is that *same* responses are benefited more by sequential presentation than are *different* responses. This suggests that there is a facilitation in processing the second of two identical patterns presented sequentially that is not a factor when the patterns are presented simultaneously.

This model, however, is difficult to reconcile with our data. On the one hand our experiment did not involve any naming process. Secondly, although we cannot exclude a priming effect (stimuli in each pair are identical except for the timing of one or two events), this effect cannot explain the reaction time differences between our *same* and *different* responses for the following reason. All stimulus pairs were presented twice, once for the vocal and once for the clapstick rhythm judgment. The two presentations of a stimulus pair were physically identical, and hence, if there was a priming effect (due to the similarity of the stimuli), it should be the same for the two presentations. However, for *sd* (vocal rhythms=same, clapstick rhythms=different) and for *ds* pairs (clapstick rhythms=same, vocal rhythms=different) we observed huge reaction time differences between *same* and *different* judgments. Clearly, these differences were not caused by stimulus differences (they are identical) and ensuing different priming effects, but had to be attributed to different processes involved in the same-different judgment.

Our results are compatible with the hypothesis that *same* and *different* responses are based on two different processes (e.g., Bamber, 1969; Beller, 1970; Keuss, 1977; Silverman & Goldberg, 1975), and also with the structure-mapping model of Markman and Gentner (2005). *Same* responses are generally attributed to a fast, holistic matching
process, whereas different responses are attributed to a slower, analytic process. However, this slower process does not necessarily imply identification of differences, which is not required in our experiment. It could simply provide, as suggested by the Markman and Gentner model, a specific alignment of a common structure and some alignable differences that lead to the different judgment. An alternative hypothesis is that reaction time differences for same-different judgments can be attributed to a single matching process. According to this hypothesis (e.g., Briggs & Johnsen, 1973), the difference in reaction times arises because of rechecking that is performed for different judgments. The most likely reason why rechecking might be necessary for different judgments is that internal noise is more likely to result in spurious mismatches of features than in spurious matches (Krueger, 1978).

The second, most remarkable observation to address here is the fact that clapstick and vocal rhythm reaction times are similar in the different-condition, but significantly different in the same-condition. This suggested that the two types of rhythm undergo different types of processing. In the same-condition judgments on vocal rhythms had a notable advantage of 195 ms over clapstick rhythm judgments, that is, vocal rhythms seem to be processed with preference. In the different-condition this preference seems to be obscured by the greater difficulties of determining differences for vocal rhythms. This is clearly indicated by the error analysis: there were fewer errors for vocal rhythms in the same condition, but considerably more in the different condition than for clapstick rhythms. Vocal rhythms seem to have a processing advantage over clapstick rhythms, but specifying their features (in the different
judgment) appeared to be more complex and error-prone than those for clapstick rhythms.

A possible explanation for the higher error rate for vocal rhythms in the different-condition may be that, in comparison to clapstick rhythms, the change of vocal rhythms was subtler and more complicated to process. One of the features of the clapstick sounds was, especially in this experiment, that the spectrum of clapstick stimuli was more stable than that of the vocal stimuli: the clapstick sound used in this experiment changed neither pitch nor timbre, whereas the vocal had large variations and spectral changes. As described in Chapter 4, various sound features contribute to form rhythmic patterns, e.g. pitch, intensity, spectral dynamics and timbral changes, duration, etc. All these elements alter continually to form the rhythmic patterns of the voice. However, the clapstick sounds in this experiment were all identical. In other words, the only element that contributed to the clapstick rhythmic pattern formation was the temporal relationship between unchanging spectral events. Therefore, it may be easier to make decisions on clapstick than on vocal rhythm.

An additional element that may contribute to the greater complexity of vocal rhythms may be the fact that they showed more events per time units than the clapstick rhythms. In our materials the average of vocal intervals per excerpt was 6.73, and the average of clapstick intervals was 4.35 per excerpt.
The significance of these factors, however, has to be explored in future research that will probably use other types of instrumental sound than clapsticks, perhaps some pitched percussion instruments or even melodic instruments.

6.7 Summary of behavioral experiments

The design of the experiments was such that participants listened to identical sets of sound pairs during both the vocal and the instrumental rhythm task. Therefore differences in behavioral responses are not due to different stimuli but must be attributed to differences in the tasks and/or the task associated component rhythms (i.e. vocal and instrumental). In addition, as none of our participants understood the language of the vocal stimuli, we can also exclude any linguistic effects on the behavioral measures of this experiment.

Interestingly, during the post-experimental debriefing, most participants (10 out of 14) reported that clapstick rhythms were much easier to hear and recognize. However, the results show a more complex situation: in the same condition reaction times were significantly longer and there were more errors for the clapstick than for the vocal decision; i.e. the fact that participants thought decisions on clapstick rhythms were easier, did not show in either their reaction times or in the error rates for the same condition. On the other hand, in the different condition reaction times were shorter and error rates lower for the clapstick than for the vocal decision. Another difference was that for clapstick decisions the error rate did not vary significantly across the four conditions, whereas in vocal tasks the error rate increased significantly from the ss
condition, where it was lowest, to the dd condition with the highest value. This error distribution suggests that the vocal rhythm task was influenced by the task-irrelevant rhythm, whereas the clapstick rhythm task was not. Though the fact that participants subjectively perceived clapstick rhythms as easier may indicate perceptual differences between the two types of rhythms (e.g. higher event density of the vocal rhythms, fixed spectral dynamics of the clapstick rhythms, etc.), also may also reflect an influence of western music education, which centers more on instrument music than vocal music; consequently, participants may be more familiar with determining rhythms of instrumental sounds.

In conclusion, the analysis of the behavioral results showed that processing of vocal and clapstick rhythms was different in terms of reaction time as well as error rate. We interpret these results as indicating at least two ways of rhythm processing that may reflect either (partially) different pathways or different processing strategies for sounds with vocal and instrumental timbre. To address this latter possibility, we are going to perform fMRI experiments using the same set of stimuli. We expect that these experiments will help us to a better understand where and how the brain processes these two types of rhythm.
7 fMRI passive listening experiment

7.1 Introduction to fMRI experiments

The main behavioral experiment shows that the reaction times vary depending on the rhythm conditions. The differences in reaction time indicated participants were processing vocal and instrumental rhythms differently. However, one limitation of behavioral studies is that they cannot tell us how and where the brain processes the rhythmic information. If we want to know more about the localization of the brain activities, other type of experiments are necessary.

Previous brain studies on rhythmical or temporal information (e.g. Grahn and Brett, 2007, Zatorre 1998, Sakai, 1999, Peretz and Kolinsky, 1993) have mostly been performed by using electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI). These methods register various brain activities and they permit to answer different questions than behavioral studies. Yet, each method has its advantages and disadvantages. Depending on the nature of research questions, different methods may be applied. For example, EEG and MEG are excellent with regards to the time resolution--- the brain response to the stimulus can be recorded almost simultaneously. Many event related potential (ERP) and field (ERF) studies were done by EEG and MEG for this good reason. However, these two methods
are not very good at localization and usually we can only get a very rough idea about ‘where’ in the brain the observed activations take place. Even though the MEG may have better spatial resolution than EEG, MEG is still does not detect all the activities as fMRI does.

fMRI measures the flow of oxygenated blood or the blood oxygenated level dependent effect. Roy and Sherrington (1890) showed that a particular area of blood flow increased when that areas’ neurons are firing. More neuronal activities require more glucose and oxygen to be delivered through the blood stream, so the oxygen level is thought to be an indicator of neuronal activity in the brain (Logothetis, 2001). The measurements of fMRI are not the neuronal activities directly, but the blood oxygen level changes. Therefore, the ‘activation’ we obtained, i.e. the changes in blood oxygen levels, indicates locations (where) and levels (how) of brain responses during the experimental task.

On the other hand, fMRI is known for its good resolution of localization. In addition, fMRI not only presents where the activations take place but also the level of activations. So we would know the level of significant difference of activation. However, fMRI may not be good to present temporal information, but temporal resolution is not important in this experiment. The present study focuses on whether vocal and instrumental timbre influence rhythm perception, not the rhythm information itself. The research question is ‘where’ and ‘how’ the brain responses to the stimuli and the temporal resolution are less relevant.
7.1.1 Use of Scans

One of the main concerns of auditory fMRI studies is that the scanner noise may interfere with the perceptibility of the experiment stimuli and the participants may not be able to focus on the presentation. Depending on the scan design and the scanner, the noise can be from 100 dB to 130dB. Especially as we are investigating the rhythm processing, scanner noise produces some regular and repetitive rhythm and may confuse and impede the perception our participants.

In a pilot study, we first used a continue scan: the scanner was on the whole time. However, the problem with the continue scan was that the amplitude and frequencies of the scanner noise were such that one could not hear the stimuli well as the scanning process was going on. We noticed that the instrumental sounds were masked by the scanner noise even when the vocal sounds were still clearly heard. Therefore we modified the timbre and frequency of instrumental sound to make sure that both rhythms can be heard without difficulties when scanner noise occurs during the experiments. However, even though we had improved the audibility, and did manage to get a more audible timbre, the instrumental sound was still not clear enough for the decision task.

We then considered different scan procedures to reduce the scanner noise, sparse sampling and clustered acquisition. The sparse acquisition, as it was e.g. used by Sakai et al (1999), is that the participants would not hear any scanner noise during the sounds presentation and decision making. In their study, the timing of the scans starts with a 4
seconds rhythm presentation, followed by 9 seconds memorization (delay) period and 1.8 seconds scan, then finally 5.2 seconds for the reproduction of the stimuli. The total scan time for one stimulus is 20 seconds. The brain activation would only reflect the retention of rhythms because no stimulus occurs during the last 9 seconds prior to the scan. However, the disadvantage of this type of scan is that there are only 1.8 seconds of scanning time and too few volumes per trials can be analyzed from this paradigm.

As we were not sure what point during each trial they make comparison, we need a longer scanning time and not miss out some brain activities. In particular, in our experiment, participants would hear and memorize a first stimulus (ca. 2 seconds) and then a second stimulus (ca. 2 seconds) with an inter-stimulus interval of ca. 1 second. They would then compare these two stimuli and making decision on whether the two rhythms of these stimuli are the same or not. So that they need memory, comparison, and decision making from the sound presentation starts until the end of one task, which takes about 7 seconds. We are interested in the activations when they are making comparison between two stimuli, which is the beginning of second stimulus and last until end of second stimulus. During these 7 seconds, we expect the Blood Oxygenation Level Dependent (BOLD) activation starts from the beginning the presentation and reaches its maximum around 7 to 9 seconds.

We then considered another type of scan procedure: clustered scans. This is the mode of scanning in which the acquisition of the scans is concentrated in the clustered part of the acquisition time while the remaining of repetition time the scanner is not
activated. In our design, we have 4 seconds repetition time the scanner was acquiring images during the first two seconds while it was off for the rest of repetition time. So the scanner is on and off every other 2 seconds. Half of the scan starts with the first sound file of each pair; the other half starts with scanner off. With clustered scans, we can then observe the activation of whole experiment. Yet, these types of scans produce noise during the tone presentation and could interfere participants’ listening. However, the noise level is much less than the block scans and with earplug and noise reduction headphones, participants were able to focus on the task during the scans. Therefore, we decided to use clustered scan for the experiment.

In addition to the clustered scans, we also performed two runs of block scans. These block scans were performed as passive listening. Three types of presentations were included in block scans: voice only, instrumental sounds only, and voice and instrumental sound combined. The purpose of these scans was to distinguish how the brain responds to each type of stimulus without any task. In the clustered scans, participants have to listen, compare, and process stimuli. Therefore, the activation we obtain that includes multiple processes. The block scans allow us to determine brain activations due to listening only, without assigned tasks, and activations caused by the vocal and instrumental stimuli separately, and we would be able to tell whether listening to these different stimuli produces different activation patterns. In addition, comparing the activation patterns of the two types of scans, may allow us to sort out which of activations is due to listening, comparing, or decision making.
After we decided to use clustered scans for our main experiment and block scans for passive listening task, we had another test run in order to check that our participants will have no difficulties perceiving the stimuli during the experiments.

7.1.2 Participants

Six English-speaking musicians participated in this experiment. In addition to the recruitment criterion of behavioral experiment, the participants should not have any neurological or psychiatric disorder, should not pregnant, no metal implants, and have to be right-handed.

Participants had done the handedness test that was published on web by the Brain Mapping center at UCLA, which was designed by Oldfield (1971). The participants included 3 females and 3 males (mean age = 30.83 years) All were graduate music major students in the School of Music at Ohio State University. Their area of concentration included performance and musicology. The average age at which the musicians first received their music training was 8. The average years of music training is 21.17 years. All of them speak English as their native language. All of them reported without any hearing, speech, language, and neurological or psychiatric disorder. They received a compensation of 25 US dollars for their participation.

7.1.3 Procedure

Before the experiments, all participants were screened by the fMRI lab research scientist, to insure the safety of participating in the experiment, e.g. no metal implant,
no history of mental disorder, etc. Then participants receive a brief instruction about the experimental task and performed 10 practice trails prior the experiments.

Stimuli were presented over electrostatic headphones inserted into noise-reduction ear defenders. Further deduction of scanner noise was achieved with insert earplugs. When wearing earplugs and headphones, participants reported the stimuli could be heard clearly. Participants were instructed not to move any part of their body during the scan, except right index and middle finger for the button pressing.

7.1.4 Stimuli

Materials are identical to the stimuli in behavior studies.

7.1.5 Scanning sequence

Each participant would be scanned for 1 anatomic scan, 2 B0 maps (filed map), 4 runs of clustered scan, and 2 block scans. Order of presentation of the 4 runs of clustered and 2 blocks was counterbalanced across participants. Half of our participants started with the anatomic scan, B0 maps, clustered scan, and finally block scans. The other half of the participants started with the anatomic scan, B0 maps, block scans, and then clustered.

7.1.6 fMRI data acquisition

Images were acquired using a Philips Achieva 3T MR scanner (Philips Medical System, Andover, MA, USA) equipped with an 8-channel receive-only radio frequency (RF) coil (Philips). The body RF coil was used for transition. High resolution T1-
weighted anatomical volume images data were acquired with the following parameters:
TR/TE 8.1/3.7 ms, 512×512 matrix, 0.43x0.43x1.00 mm³ voxels, 160 slices, 3D axial
acquisition, 1.5 SENSE acceleration factor, and 7.5 minutes total scan time. T2*-weighted functional images were acquired using a blipped single-shot gradient-echo
EPI imaging sequence with an isotropic spatial resolution of 3.4 mm. The acquisition
parameters for the two passive listening runs were: TR/TE 2000/22 ms, 80° flip angle,
64 × 64 matrix size, 38 slices, and echo train of 35 ms. with blood oxygenation level
dependent (BOLD) contrast. In addition, an image set with similar resolution to the
functional scan and with two gradient recalled echoes was used to measure the B0
inhomogeneity map. The time between the two echoes is 1.4 ms.

Functional data were acquired in two runs of 170 volumes each for the passive
listening task and four runs of 106 volumes each for the main task. Volumes within runs
were acquired continuously in an ascending interleaved order. The first five volumes of
each session were discarded to allow equilibration of tissue magnetization.

Stimulus presentation was accomplished using an Invivo Eloquence system
(Invivo International) equipped with a 15 inch diagonal LCD monitor with a 1024 X
768 resolution that displayed the task instruction at the beginning of each run. The
system provided the auditory presentation using electrostatic stereo headphones with
passive noise reduction. The timing of the presentation was controlled by E-Prime
software (Psychology Software Tools, Inc.) and it is synchronized with the MRI
scanner.
7.1.7 Individual and group activation maps

The imaging data were analyzed using FSL (Smith, Jenkinson, Woolrich, Beckmann, & Behrens, 2004) software. Pre-processing consisted of motion correction with FLIRT (Jenkinson, Bannister, Brady, & Smith, 2002), B0 unwarping (with effective EPI echo spacing of 0.55ms, EPI TE of 22 ms, and signal loss threshold set to 10%), slice timing correction, BET brain extraction (Smith, 2002), spatial smoothing with a FWHM 6 mm Gaussian kernel, and high-pass temporal filter cutoff of 50 sec. In order to identify the areas of the brain activated by different tasks, individual and group analyses were performed. The BOLD response was modeled by convolving the neural functions with a canonical hemodynamic response function (HRF) and its temporal and dispersion derivatives to yield regressors in a General Linear Model (GLM) that modeled the BOLD response to each event type. For first level analysis we used FILM prewhitening, added motion parameters to the model to create individual activation maps, and then, because of the small group size, both fixed effect and mixed effect analyses (FLAME 1) were used to generate average group maps. Z statistic images were obtained using cluster analysis with clusters determined by Z >3.1 and a cluster significance threshold of p = 0.05 (Worsley, 2001). All functional images were registered to the subjects’ structural images, which in turn were registered to the Montreal Neurological Institute (MNI152_T1_2mm) standard brain.

The assumption of normal data distribution that is underlying the mixed effect analysis is questionable due to the relatively small number of participants in our study. Therefore, we base our analysis on the fixed effect analysis. However, as the mixed
effect analyses gave similar results, we have good reason to assume that the results from our analyses are generalizable beyond the actual sample of six subjects.

### 7.2 Stimulus sequence

Four different types of stimuli were used in two runs. The first run included pairs of clapstick and vocal sound, and the second run included pairs of combined sounds (clapstick and voice) and clapstick. The clapstick pairs of the first and second run were acoustically identical, but for those in the second run the beat of was continuing throughout the stimulus presentation, i.e. the clapstick stimuli of the two runs differed only in the temporal alignment of their two segments. Unfortunately we cannot compare the two runs directly because of incompatibilities between the two runs. As interesting as this comparison may be, we must leave the comparison in between the aligned and non-aligned rhythms to be addressed by a future experiment.

Each run consists of 17 blocks and each block is 20 seconds long containing 7 stimuli. Each stimulus presentation took 3 seconds, except for the last one, which is only 2 seconds long. The presentation order for the 20 blocks in the first run was: abacabacabacabaca (a = rest period, b = instrumental sounds only, c = vocal sounds only) and the presentation order for the 20 blocks in the second run was: adaeadaeadaeadaea (a = rest period, d = both sounds combined, e = claps with aligned beat).
7.3 Task procedure

During the passive scans, participants were asked to relax and not to pay particular attention to the stimuli. Neither did they have to make decisions or press button.

7.4 Data analysis

The scans of first participant served as a pilot study. We used the same stimuli as for the other participants but with a different scan sequence. After the first scan we realized that it took too much time and participants may get tired during the scan. So we then redesigned the scanning sequences for the following scans. Due to the incompatibilities of the different scanning sequences the scans from the first subject had to be excluded from the passive task analysis.
We performed a statistical parametrical analysis of the functional images from the five participants to identify the brain areas associated with the (passive) processing of the stimuli. First data for each participant was analyzed separately, and then a group analysis was performed for each of the two runs separately. Although the statistical analysis of Run 1 and Run 2 had to be done separately, mainly because subject positions changed slightly and there were different movement artifacts between the two runs, the separate scans were nevertheless comparable, because they were taken under the same condition, participants, almost the same time (one followed by the other). In addition, they were based on demeaned activity levels.

7.5 Results

7.5.1 Stimulation with clapstick rhythms

Fig. 21 shows both right and left hemispheres were activated. The regions that showed statistically significant activation included: bilateral Planum Temporale, bilateral Superior Temporal Gyrus (posterior division), left Heschl’s Gyrus (includes H1 and H2), left Supramariginal Gyrus (posterior division), right Supramariginal Gyrus (anterior division), right Middle Temporal Gyrus (temporooccipital part), right Inferior Frontal Gyrus (pars triangularis), right Frontal Pole, and right Frontal Orbital Cortex. In addition, the Cerebellum was also activated.
Figure 21: Average activations by clapstick stimuli. Right hemisphere is displayed on the left; anterior is up.

7.5.2 Stimulation with vocal rhythms

The activation regions that were statistically significant included: bilateral Superior Temporal Gyrus (posterior division), bilateral Planum Temporale, left Heschl’s Gyrus (included H1 and H2), left Temporal Pole, right Precentral Gyrus, right Postcentral Gyrus, and Cerebellum.
Figure 22: Average activations by vocal stimuli. Right hemisphere is displayed on the left; anterior is up.
7.5.3 Comparison between two stimuli

In general, activation areas that obtained significant z-value for both clapstick and vocal were similar (i.e., Superior Temporal Gyrus (posterior division), Planum Temporale, and Heschl’s Gyrus (included H1 and H2). These three areas are known to be the sound sensitivity areas and they are all part of the STG.

Figure 23: Axial (left) and coronal (right) sections through the main activation areas in the passive listening task. The activation areas of clapstick (Z>3.2, yellow-red) and voice (Z>3.2, blue). When overlapping, the colors turn greenish. Right hemisphere is displayed on the left; anterior is up in the left image.
There were some activation areas that only showed up for clapstick sounds, for example, right inferior frontal gyrus, right frontal pole, right insular cortex, left cingulate gyrus (posterior), precentral gyrus.

The right frontal pole was the only region significantly activated when we test for regions more activated by clapstick then vocal stimuli.

Some activations only appear in vocal rhythm, for example, temporal pole, frontal orbital cortex, postcentral gyrus. A statistical comparison of vocal vs. clapstick showed that Planum temporale, Heschl’s gyrus, STG posterior, and central opercular cortex had significant higher activation bilaterally for vocal than for clapstick.

However, when we compared the maximum activation for the two stimulus types, the maxima were not overlapping, even through the two areas were very close within the posterior division of the right Superior Temporal Gyrus (STG): the vocal maximum was clearly anterior to the clapstick maximum (see fig. 24). The highest activation for the two stimuli in the left STG was not as clearly separate as in the right side. Although the highest intensities of two sides were very close, the activation patterns were very different.
Figure 24: Axial (left) and sagittal (right) sections through highest activation areas. The displayed maximum activation for the clapstick sounds is $Z>9$ (yellow-red) and for the voice sounds it is $Z>14$ (blue).
7.5.4 Stimuli of combined sounds

The regions that were highly significant included: bilateral Planum Temporale, bilateral Superior Temporal Gyrus (posterior division), left Heschl's Gyrus (includes H1 and H2), and left Middle Temporal Gyrus (temporooccipital part). In addition, other regions with lower z-value (still significant) included: left Supramarginal Gyrus.
(posterior division), right Lateral Occipital Cortex (superior division), right Supramarginal Gyrus (anterior division), right Postcentral Gyrus, right Angular Gyrus, and left Superior Parietal Lobule.

Planum Temporale, Superior Temporal Gyrus (posterior division), and Heschl's Gyrus (includes H1 and H2) again had highly significant activations for this stimulus type.

7.5.5  **Stimuli of clapstick (the pulses of the two stimuli aligned)**

The regions these were statistically significant included: bilateral Superior Temporal Gyrus (posterior division), bilateral Supramarginal Gyrus (posterior division), bilateral Planum Temporale, and left Heschl's Gyrus (includes H1 and H2).

<table>
<thead>
<tr>
<th>Activation areas</th>
<th>Clapstick (rhythm aligned)</th>
<th>Voice</th>
<th>Combined</th>
<th>Clapstick (rhythm aligned)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left Planum Temporale</td>
<td>12.1 (-60, -30, 10)</td>
<td>14.1  (-52, -22, 4)</td>
<td>13.9 (-52, -20, 2)</td>
<td>9.99 (-54, -32, 8)</td>
</tr>
<tr>
<td>Right Planum Temporale</td>
<td>7.29 (-50, 26, 8)</td>
<td>16.9  (-66, -18, 6)</td>
<td>18.3 (-62, -18, 6)</td>
<td>8.42 (-58, -12, 2)</td>
</tr>
<tr>
<td>Left Superior Temporal Gyrus, posterior division</td>
<td>9.4 (-64, -22, 6)</td>
<td>15.5  (-64, -28, 8)</td>
<td>14.1 (-64, -26, 8)</td>
<td>5.87 (-66, -18, 4)</td>
</tr>
<tr>
<td>Anatomy</td>
<td>t-stat</td>
<td>x, y, z</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>--------</td>
<td>---------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Superior Temporal Gyrus, posterior division</td>
<td>12.1</td>
<td>(64, -32, 6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16.9</td>
<td>(62, -20, 6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13.2</td>
<td>(66, -28, 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9.81</td>
<td>(66, -32, 4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Superior Temporal Gyrus, anterior division</td>
<td>14.9</td>
<td>(62, 0, -2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Heschl's Gyrus (includes H1 and H2)</td>
<td>8.63</td>
<td>(-52, -10, 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16.5</td>
<td>(-54, -8, 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16.4</td>
<td>(-54, -10, 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.36</td>
<td>(-50, -14, 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Supramarginal Gyrus, posterior division</td>
<td>6.89</td>
<td>(-64, -50, 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.54</td>
<td>(-62, -50, 10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.28</td>
<td>(-64, -48, 10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Supramarginal Gyrus, posterior division</td>
<td>7.5</td>
<td>(70, -16, 14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.26</td>
<td>(64, -18, 42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Supramarginal Gyrus, anterior division</td>
<td>4.23</td>
<td>(42, -56, 52)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Middle Temporal Gyrus, temporooccipital part</td>
<td>6.89</td>
<td>(68, -44, -12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Middle Temporal Gyrus, temporooccipital part</td>
<td>8.54</td>
<td>(-62, -50, 10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Angular Gyrus</td>
<td>6.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brain Area</td>
<td>Z-Value</td>
<td>Coordinates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------------</td>
<td>---------</td>
<td>---------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Inferior Frontal Gyrus, pars triangularis</td>
<td>5.88</td>
<td>(54, 32, -2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Frontal Pole</td>
<td>5.49</td>
<td>(40, 50, 6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Postcentral Gyrus</td>
<td>5.26</td>
<td>(60, -16, 46)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(64, -18, 42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Lateral Occipital Cortex, superoior division</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(36, -64, 54)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Frontal Orbital Cortex</td>
<td>4.16</td>
<td>(30, 28, 4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Precentral Gyrus</td>
<td>5.32</td>
<td>(-52, -2, 48)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Superior Parietal Lobule</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-34, -56, 62)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Middle Frontal Gyrus</td>
<td>3.84</td>
<td>(-44, 6, 52)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2**: Activation areas for all stimuli in passive fMRI scans, showing z-values, and coordinates of maximum activations.
The dominating view in neuroscience concerning the relationship between language and music processing by the brain is that they are processed by different neuronal networks (Adams and Janata, 2002; Levitin & Menod, 2003; Koelsch et al., 2002; Tillman et al., 2003; Hebert et al 2003). Research (Bent et al., 2006, Levy et al., 2001, Vouloumanos, 2001, & Zatorre et al., 2002) has shown that humans show different cognitive processing of speech and non-speech sounds. Some studies show a left hemispheric dominance for linguistic patterns whereas the right hemisphere was found to dominate in music processing. However, the activations for clapstick and vocal stimuli in the present study are clearly bilateral. Our finding is similar to those of Zatorre et al. (2001) showing that both speech and music are processed bilaterally. In
addition, Belin (2000) showed that vocal stimuli usually produce higher activation than non-vocal stimuli. This is comparable to our findings that vocal rhythms produce higher activations than clapstick rhythms.

In the present study the overall activations for the three different stimuli were similar in this passive listening task, though with some variations. Both right and left auditory cortexes were activated during vocal and instrumental sound presentation. This was in agreement with the findings of Schoen et al (2005) that music and language are processed by similar and overlapping brain areas despite the fact that physical features of the sounds utilized were different.

Vouloumanos et al, (2001) showed greater activation by speech stimuli than non-speech. The present study found higher activation in vocal than clapstick. The higher activation for voice stimuli in our study cannot be due to any lexical effect because the lyrics were in a language no participant understood.

Interestingly, although the activated regions that reached significance were very similar, areas of highest activation for clapstick and vocal were clearly separated on the right hemisphere, both within STG posterior, vocal in an anterior and clapstick in a posterior position. This indicates that there were groups of neurons that either responded specifically or with higher intensity to one of the two stimuli in the passive listening task. So, besides networks of neurons that were activated by both types of stimuli, there were also groups of neurons specifically activated by only one of them. This suggests that some groups of neurons in the right posterior STG perform a kind of
acoustic object identification, identifying stimuli in terms of spectral content (i.e. voice vs. clapstick) and, possibly, temporal features (indicated by the fact that these neuronal clusters were found in the right hemisphere).

In addition, the highest activation patterns on the left hemisphere were very different than the right. The activation regions were much smaller on left than on the right; images showed many small spots, some were activated by only one stimulus type, others by both. The difference between the activations in the right and left hemisphere can probably be explained in terms of their assumed functional differences.

Previous studies (e.g., Geschwind and Galaburda, 1987) have shown that right hemisphere matures earlier than the left, and the right STG may already become functional during fetal development (Chi, Dooling, and Gilles, 1977). The right hemisphere is suggested to process at a lower temporal resolution (e.g. rhythm perception) and to perform a more global analysis (Ivry and Robertson, 1998).

Therefore, our results for the posterior STG activation could reflect both the different sound objects (vocal/clapstick) as well as their different rhythmic structure.
8 fMRI decision task

8.1 Stimulus sequence

The same 40 stimuli pairs as in the behavioral experiment were used, with slight modification in the timing of presentations to accommodate for the scan sequence requirements: for each sound pair presentation, the first stimulus was ca. 2 seconds, followed by an Inter-Stimulus-Interval (ISI, from the end of first stimulus to the onset of second stimulus) of ca. 1 second, and then a second stimulus.

After the sound presentation, there was a 2 seconds decision period. After the decision period, there was a 0.1 second 1 KHz beep. Participants were asked to press a button after the beep in order to register their answers. The decision period was followed by a 11 seconds rest period, which ended one stimulus presentation cycle.
8.2 Task procedure

Participants were the same as for the fMRI passive experiment. Participants were instructed to pay attention to the screen for the task announcements before sound presentations. They were not allowed to move any body parts (except index and middle fingers of right hand for behavioral responses) throughout the experiment. The sound was delivered to a headphone via a pair of plastic tubes.

The screen would show either “is vocal rhythm the same?” or “is clap rhythm the same?” 10 seconds before the task changed. No image (screen turn black) was presented during the whole stimulus presentation. They then listened to one pair of sounds and decided if the rhythm was the same or not. They were instructed to hold
their decision until they heard the beep. Participants’ responses of decision making were recorded via a response box. They pressed the right button with the right index when the rhythms were the same or the left button with the right middle finger when the rhythms were different. The button presses always occurred at the beginning of the 11 seconds rest period, and the rest period is always 11 seconds regardless of when exactly the buttons were pressed.

**Figure 28: The task procedure.**

4 runs were presented in the whole experiment; 2 blocks were presented in each run; each block contained 10 stimulus presentations. One stimulus presentation was 18 seconds; so one block was 180 seconds and two blocks were 360 seconds; plus a 30
seconds dummy scan before each run starts; so the total is 390 seconds (about 6 minutes scanning time) per run. 4 runs are about 24 minutes scanning time.

Figure 29: Stimulus sequence.
8.3 Data analysis

We performed a statistical parametrical analysis of the functional images from the six participants to identify the brain areas involved in the task performance. Data for each participant was first analyzed separately, then a group analysis was performed for the four runs together. Two types of task were analyzed separately.

The following images show the brain activations that reached the significance level of the z value of 3.7 in the two tasks.
Figure 30: Average activations of clapstick rhythm decision at different slice. Right hemisphere is displayed on the left; anterior is up.
Figure 31: Average activations of vocal rhythm decision at different slice. Right hemisphere is displayed on the left; anterior is up.
**Figure 32**: Coronal sections showing the main activation areas for clapstick and vocal rhythm decision task. Left (red): clapstick rhythm task. Right (yellow): vocal rhythm task.

<table>
<thead>
<tr>
<th>Activation areas</th>
<th>Clapstick</th>
<th>Voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right Superior Temporal Gyrus, posterior</td>
<td>25.5 (64, -22, 6)</td>
<td>26 (64, -22, 6)</td>
</tr>
<tr>
<td>Left Planum Temporale</td>
<td>27.5 (-52, -22, 4)</td>
<td>29.2 (-52, -22, 4)</td>
</tr>
<tr>
<td>Right Heschl’s Gyrus</td>
<td>23.1 (52, -18, 4)</td>
<td>24.4 (52, 18, 4)</td>
</tr>
<tr>
<td>Left Heschl’s Gyrus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Supramarginal Gyrus, posterior</td>
<td>19.2 (-60, -48, 10)</td>
<td>20 (-60, -48, 10)</td>
</tr>
<tr>
<td>Activation Area</td>
<td>z-value</td>
<td>Coordinates</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>---------</td>
<td>--------------</td>
</tr>
<tr>
<td>Right Central Opercular Cortex</td>
<td>15.4</td>
<td>(62, -18, 18)</td>
</tr>
<tr>
<td></td>
<td>15.1</td>
<td>(62, -18, 18)</td>
</tr>
<tr>
<td>Right Precentral Gyrus</td>
<td></td>
<td>12.2 (58, -4, 40)</td>
</tr>
<tr>
<td>Left Putamen</td>
<td>6.81</td>
<td>(-18, 14, 0)</td>
</tr>
<tr>
<td>Right Frontal Orbital Cortex</td>
<td></td>
<td>11.7 (32, 28, 4)</td>
</tr>
<tr>
<td>Left Frontal Orbital Cortex</td>
<td>10.07</td>
<td>(-32, 28, 0)</td>
</tr>
<tr>
<td>Left Thalamus</td>
<td>5.34</td>
<td>(-12, -16, 6)</td>
</tr>
<tr>
<td>Left Cingulated gyrus, posterior</td>
<td>5.74</td>
<td>(-8, -26, 40)</td>
</tr>
<tr>
<td>Right Lingual gyrus</td>
<td></td>
<td>5.26 (4, -78, 2)</td>
</tr>
<tr>
<td>Right Intracalcarine Cortex</td>
<td></td>
<td>4.01 (10, -84, 10)</td>
</tr>
</tbody>
</table>

Table 3: Activation areas, z-value and coordinates of both stimuli. Listed areas are only local maxima.

The table above shows the z-value and coordinates of maximum activations for the two tasks. The activation regions in this decision experiment showed some similarities and dissimilarities with our passive listening task. First, the vocal decision task obtained higher activation (z value) than the clapstick task; this was the same as in passive task. However, unlike for passive listening, for which the highest activated
regions for the two rhythms were different, we found the highest activation region in the
decision experiment was the same for both rhythm decision tasks, the left Planum
Temporale. The coordinates of some local maxima were even exactly the same between
two tasks (i.e., Right Superior Temporal Gyrus, posterior, Left Planum Temporale,
Right Heschl’s Gyrus, Left Supramarginal Gyrus, posterior, Right Central Opercular
Cortex (Fig. 36 &37)). The left Heschl’s Gyrus had a relative high activation for vocal
but not for clapstick. This may due to the left Heschl’s Gyrus being part of the language
processing areas of the brain, and the decisions on vocal rhythms seem to activate this
area even though it contains mainly the primary auditory projections. Moreover, unlike
in the passive task, more areas were activated by vocal than by the clapstick task (fig.
25).

In order to compare differences between the two types of tasks, including those
activation areas that only appeared in one task but not the other, we performed two
types of additional image processing. First, we subtracted areas activated in the
clapstick task from those activated in vocal task. This allowed us to determine areas
only active in the vocal task. We did the reverse subtraction to determine areas only
active in the clapstick task. The following images give a representative example for
each task:
Fig. 33 confirms that the brain areas activated in the two tasks are largely overlapping. Only the vocal task produces some specific activations in left frontal lobe and some subcortical areas. However these specific activations are all marginal zones of the common, overlapping areas. This suggests the vocal task may activate slightly larger neuronal populations than the clapstick task.

However, interesting task related differences were found when comparing the overlapping areas in more detail. For this we performed another image processing that permitted us to determine where in the overlapping areas one of the task produces a stronger activation than the other. We obtained the following results:
Figure 34: Coronal sections showing areas in which either the clapstick task (left; red) or the vocal task (right; yellow) produces higher activations.

Fig. 34 showed that some areas had stronger activation in clap, and some in vocal. Over all, the comparison of the two images clearly showed that the vocal task elicited higher activation level than the clapstick task in several regions. Activations for the clapstick task were dominant in the right supramarginal gyrus posterior, right inferior frontal gyrus, left parietal operculum cortex, and left precentral gyrus. Activations for the vocal task were dominant in the right planum temporale, right central opercular cortex, right STG posterior, right middle temporal gyrus posterior, right postcentral gyrus, left precentral gyrus, left planum temporale, left STG anterior, left Heschl’s gyrus, and left STG posterior.

In order to see more details in selected regions, a region of interest (ROI) analysis was conducted to test difference in activation of specific areas. These areas
included: Left caudate, right caudate, left putamen, right putamen, left superior temporal gyrus (posterior), right superior temporal gyrus (STG) (posterior), left superior temporal gyrus (anterior), and right superior temporal gyrus (anterior).

We performed t-tests on the task-related mean signal changes of the ROIs. The results show that, except for the right putamen, the vocal task always obtained higher value than clap task. Notably, the STG posterior and STG anterior had much higher value than other regions. The activation of the STG posterior and STG anterior were bilateral, but the extent of differential activation was greater in the left hemisphere. Clap and vocal rhythm tasks lead to different signal changes in the left anterior and posterior STG and right anterior STG. The right caudate also showed considerable difference, but its p value was not significant. This may be due to the fact that, despite the substantial mean difference, there seems to be a considerable variation across subjects, which, together with the small number of subjects (n=6) prevents the mean difference to reach significance. However, this can only be tested by a repetition of the experiment with a larger number of subjects.
Figure 35: Signal changes (in %) for clapstick and vocal rhythms tasks in the regions of interest (ROI).

<table>
<thead>
<tr>
<th>Region</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L Caudate</td>
<td>-0.872</td>
<td>0.4233</td>
</tr>
<tr>
<td>L Putamen</td>
<td>-1.009</td>
<td>0.3594</td>
</tr>
<tr>
<td>L STG posterior</td>
<td>-4.440</td>
<td>0.0068</td>
</tr>
<tr>
<td>R Caudate</td>
<td>-1.144</td>
<td>0.3046</td>
</tr>
<tr>
<td>R Putamen</td>
<td>-0.123</td>
<td>0.9071</td>
</tr>
</tbody>
</table>
Table 4: T-tests and significances of the differences in % signal change between the vocal and the clapstick rhythm task.

<table>
<thead>
<tr>
<th></th>
<th>T-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R STG posterior</td>
<td>-1.593</td>
<td>0.1720</td>
</tr>
<tr>
<td>R STG anterior</td>
<td>-2.826</td>
<td>0.0368</td>
</tr>
<tr>
<td>L STG anterior</td>
<td>-2.665</td>
<td>0.0446</td>
</tr>
</tbody>
</table>

Figure 36: Location of Angular gyrus, supramarginal gyrus, and STG.
8.4 Discussion

Activities for vocal and clapstick tasks that reached significance level \( z=3.2 \) were largely overlapping. Even at the highest activation level were still very similar between two types of tasks. However, a region of interest analysis showed significant activation differences between the tasks in bilateral anterior STG and posterior STG.
Grahn and Brett (2007) observed that all rhythms-rest contrasts had maxima in the pre-SMA/SMA, PMd, basal ganglia, cerebellum, superior temporal gyrus, and ventrolateral prefrontal cortex/insula (all bilaterally). Local maxima for metrical simple, metrical complex and nonmetrical contrast were found bilaterally in the putamen, superior temporal gyri, left inferior frontal gyrus, and these areas may be responsible for detecting or relating the information of metrical beat intervals. In particular, anterior STG was important for the perception of regular perceptual accents. The degree of significance in STG depended on the level of rhythmic complexity. STG was activated more by the simple metrical than the complex metrical and non-metrical conditions. This conforms to the suggestion by Liegeois-Chauvel et al. (1998), that the STG may be involved in musical meter perception.

The basal ganglia, SMA, and the bilateral anterior STG were also found activated in our experiment. However, we did not find any significant activation in the insula, inferior frontal gyrus, superior frontal gyrus and amygdala.

Sakai et al., (1999) report that higher activation was obtained from integer ratio rhythm than non-integer ratio rhythm processing. However, even though our clapstick rhythms are less complex than our vocal rhythms, we obtained much higher activations for the vocal than for the instrumental task. This can be explained by assuming that the vocal rhythm task is more cognitively demanding, because in our experiment vocal rhythm were constituted by a more elaborate set of variables (e.g., spectral, intensity,
and pitch dynamics) than the clapstick rhythm. In addition, Bent et al. (2006) have shown that voice sounds leads to higher activation than non-voice sounds.

To summarize, decision tasks on vocal and clapstick rhythms produce different activation patterns. In general, the vocal task elicits higher activations than the clapstick task in all areas involved except for the right putamen; however, these differences only reached significance in the left posterior STG and in both left and right anterior STG. As stimuli were identical in both tasks, higher activation in the vocal task means that either more neurons are activated or that neurons show a higher level of activity. This can be seen as an indication that the involved areas of the STG have a preference for processing voice related information.
9 General discussion

This dissertation began with reviews of previous literature on music origins and sound production in the animal world. Fitch (2006) indicates that unlike humans, animals usually have either vocal communication or an independent manual sound production system only, and he suggests that, in order to explore the origins of music, it might be helpful not to treat music as a unitary phenomenon, but to distinguish between vocal and non-vocal practice. Then a survey of world music indicated that the roles of vocal and instrumental musics seem to be very different in many societies. In particular, it seems vocal music is more important and dominant in many cultures, and instrumental music usually serves a function of accompanying vocal music. In addition, previous studies on the processing of pitch and timbral information show that humans process pitch and timbre in vocal sounds differently than in non-vocal sounds.

Overall, there is accumulating evidence for a clear distinction between vocal and instrumental musics in the frequency domain. This brought up the question whether differences can also be found in musical rhythm processing. Hence, the subsequent and the most important part of this dissertation introduced a set of experiments to examine several aspects of human vocal and instrumental rhythms processing.
9.1 Summary and discussion of experiments

Two different experimental methodologies were used in this dissertation. The experiments were to address the question of whether there are any behavioral and physiological indications that vocal and instrumental musical rhythms are processed differently in humans. Both behavioral and fMRI experiments provided different data and allowed us to interpret them based on our research question. The reaction time data (behavioral) revealed that participants showed different reaction times, depending on stimuli (vocal or clapstick) and condition (same or different). Clapstick and vocal reaction times were similar in the different-condition, but significantly different (vocal faster than clapstick) in the same-condition. The error analysis showed that decisions on vocal rhythms might be influenced by task-irrelevant stimuli (i.e., the clapstick rhythms). While the clapstick rhythm error rates were similar across conditions (dd, ds, sd, and ss), the vocal rhythm error rates depended on conditions (with the dd condition showing the highest error rate). All these results seem to suggest that the two types of rhythm undergo different types of processing. However, our fMRI data show that this is not done in two completely separate pathways, one for vocal and the other for clapstick, because the activation regions for the two stimuli are very similar and largely overlapping.

In the passive fMRI task, both hemispheres exhibited regions with significant activations and these regions were similar for the two stimuli. Some differences showed up for the highest activation levels. The activations in this passive listening task indicate
which parts of the brain are processing the stimuli automatically, without any specific attention.

When comparing the results of the passive and the decision experiments, it turned out that basal ganglia, SMA, and bilateral anterior STG were only activated in the decision but not passive task. This indicates that these areas were involved in the required task, and, therefore, in active rhythm processing, as has also been reported by other groups (Grahn and Brett, 2007; Sardo et al., 2000; Rao et al., 1997).

In the decision tasks, the regions activated by the vocal and clapstick rhythm tasks were very similar, but the level of activation was different. In general, we found higher activation for the vocal than for the clapstick task. This finding is comparable to previous research that speech sound obtains higher attention/brain activation than non-speech sound (e.g., Bent, 2006). In particular, the differences seem only to exist at the early stage of the cortical auditory pathways, namely bilateral STG posterior and anterior. The later stage, such as the processing in the caudate nuclei and the putamen did not show any significant differences between two tasks. Previous studies (Wong et al., 2007) have shown that already early stages of auditory pathway, like the brainstem are more sensitive to vocal than non-vocal stimuli.

The fact that later or higher processing areas like the basal ganglia, do not show different activations for the two tasks suggests that the information that is processed here is more ‘abstract’, e.g. information about (abstract) rhythm, without regard to what
type of sound events actually produced these rhythms. Hence, the timbral information
does not seem to be processed at these stages.

In addition to the present study, our previous studies (Poss et al, 2008) showed
that pitch information facilitate a speech task in native Mandarin and English speakers.
In short, we used pitch contours extracted from recorded monosyllables (voice timbre)
or played on a Chinese flute as prime sounds (instrument timbre), that were followed by
Mandarin words as target sounds. The primes had no segmental information, only pitch
contours. Participants had to repeat the target words as quickly and accurately as
possible, and reaction time was the dependent experimental variable. To explore the
underlying physiological processes, the same experiments were repeated by using EEG
and fMRI methods in our lab.

One of the important findings from these studies shows that participants
(including English speakers!) show a significant facilitation when the contours of prime
and target are matched and the degree of facilitation depends on whether the prime
sound is vocal (larger facilitation) or instrumental (smaller facilitation). This indicates
that pitch information is able to influence the language processing. In other words, tonal
information and timbre of the primes do affect language processing.

Besides timbre of the stimuli, rhythmic structures are also part of sounds. A
previous study found that processing small integer ratio rhythm results in higher
activation than complex integer ratio rhythm (Sakai et al., 1999). In contrast, our fMRI
results show lower activation for both passive listening to and decisions on clapstick
rhythms than for the corresponding vocal tasks, although the clapstick rhythms contain more simple integer ratio rhythms. In addition, our behavioral experiment showed no significant effect of the complexity of the rhythm ratios on the reaction times. The results of Sakai et al. (1999) are not supported by the present study, neither by the behavioral nor by the fMRI experiment.

Overall, we suggest the cognitive processing of vocal and clapstick rhythms take place in the same or closely related pathways, but within this pathway there may be subdivisions that focus on more or less specific tasks. Nevertheless, there is a clear preference for vocal sounds: the higher level of activation for vocal sounds indicates that either more neurons are activated or neurons are more activated by vocal sounds. Furthermore, reaction times were significantly shorter for voice than for clapstick rhythms in the same condition of the behavioral experiment, indicating that people are more sensitive to the voice and can therefore respond faster.

Could it be that the way humans developed the processing of instrumental sounds is different than that of vocal sounds? When stimuli involved the human voice, higher activities are generally observed in the brain. Is that because the voice is much more important to us than other sounds? And why would that be so?

Some considerations from music archaeology, phylogeny, and ontology may help to understand why instrumental and vocal sounds may be processed differently in human cognition.
9.2 Music archaeology

In order to understand how humans process instrumental and vocal music, we can perhaps try to understand what music was and how music developed throughout human history. One way to understand music’s origin is to study music archaeology. However, archaeological works are like playing puzzles with always at least one part missing. Music itself does not fossilize; we will never know for sure how music was played in the prehistoric period or how music started in early human life. Even though some musical instruments survive and some even fossilize, they are only part of the whole story. d’Errico et al. (2003) questioned that the earliest instruments that were found, do represent humankind’s first attempts at musical instruments or music making. The earliest bone flutes (ca. 35000 years ago) that were found are already very sophisticated and there must have been some simpler instruments before them, like, for example, single-pitched flute. However, it is not easy and it may even be impossible for archaeologists to identify such flutes during excavation, because if the flute is just an un-worked tube, archaeologists may not identify it as an instrument. In addition, if musicians used perishable materials, such as leafs or wood, instruments made out of them would have disintegrated relatively quickly. So, simpler instruments as old as or older than bone flutes may simply not have survived, or, if they have, we may not be able to identify them as instruments.

On the other hand, d’ Errico et al. (2003) also mentioned that the bone flutes with several finger holes could indicate that people at that time may already have had some concepts of melodies. To develop a sophisticated instrument, a simpler instrument
must have existed before. This simpler instrument could have been a simple pipe without holes. However, another and probably the earliest ‘instrument’ humans have used were the vocal folds, the ‘instrument’ that is available to everyone at one’s birth. Then it is reasonable to assume that vocal music developed earlier than instrumental music in human history.

Other aspects of why the two types of musics developed differently come from human developmental physiology of vocal and motor learning.

9.3 Physiological connections

One of the differences between human beings and other animals is that humans have a relatively longer childhood than other animals. This means that humans learn all the skills, e.g. speaking, walking, and using hand, etc., during this prolonged childhood. On the other hand, humans do not learn all skills within the same, short period, but learn different skills at different stages of childhood (Bronfenbrenner, 1979). For example, infants can already hear sounds before birth and are able to make sounds via vocal tract at birth, but are only able to segment speech sounds when they reach the age of 9 months. Interestingly, infants are already able to recognize speech and other sounds at the 4th month. This shows the importance of speech sound to infants, even though they do not understand the meaning of speech yet but they can differentiate them from other sounds at an early stage in their life.
Figure 38: Development of auditory system before and after birth. (Carter, 2009)

The functioning of sensory organs is already established at birth, but physical and motor coordination requires a learning process. In addition, visual ability may be required for physical and motor coordination, but infants can only see about one yard at birth. So, only around the 12th month after birth, infants begin to use hands more often and to use them according to their needs.

Human ontology shows that humans acquire and develop different skills at different stages. In particular, infants are able to use their voice to express their emotion since birth, but to use hands requires more fine motor training. In addition, they are able to recognize human vs. other sounds at the very early stage of their life. Thus, humans recognize and produce vocal sounds earlier than other sounds. Because vocal sound
processing is developed at earlier stage than other sounds, the processing of vocal sound may be different from that of other sounds.

Figure 39: Development of manual dexterity and language. (Carter, 2009)

9.4 Use of music and language

The ability of vocal communication is one of the most important factors in human evolution. Paleolinguists believe good communication must be established
before any ancient society was formed in order to organize a complex work or society. Humans are the only animals that have language communication. As communication is so important and is used everyday, human must develop an ability to detect vocal sound more easily and reliably than all other sounds.

Before humans could communicate by language, they must have learned to use their voice and make some voice contours, so this ability should have arisen very early in human history. Therefore, these vocal sounds and contours could have formed part of the early human communication and music system. It is likely that there was no clear separation between language and music in the beginning.

The initial unity between language and music started to fall apart when the usage of spoken language developed and matured, and people began differentiating between language and other vocal sounds. That could be the starting point where language and music began their separate developments.

9.5 World music examples

Vocal music has the advantage feature that instrumental music does not have, that is, words and texts can be transmitted through vocal music. As language is an important means of communication in human society, language in music gives vocal music an additional dimension. Many religious songs are presented in vocal form, because text can be embedded in singing and singing elevates the text beyond the boundary of words. "There can be no doubt that originally the music of the divine service was everywhere entirely of a vocal nature (Nauman, 2010).” Examples from
Islamic call for prayer to Buddhist chanting, and shamanic chanting show the importance of singing in religious ceremonies and rituals. On the other hand, interestingly, many religious songs include text, but not all the text can be understood by humans (e.g. Shaman chanting, North American grass dance singing, some Buddhist chartings), because these non-lexical texts are sung for non-human beings (e.g. gods, sprits). Perhaps there are other reasons for why singing is so important in religious practice, for example the timbre of the human voice may attract more attention and awareness.

Not only religious songs but also lullabies are usually sung. They are only sung when people need inner harmony, because they offer relaxed, peaceful, and even magical power to heal and calm people’s minds. Some religions even eliminate instrumental music because instrumental music may distract people’s mind (e.g. Buddhism, Islam, etc.). This goes to show that vocal music provides a function that cannot be replaced by instrumental music. On the other hand, it also suggests that vocal and instrumental musics are not treated the same in our mind and body.

Many musical examples already discussed in the first chapter, show that a majority of world musics show a preference to vocal music. However, it seems that nowadays music in the Western culture focuses very much on instrumental music, although musicological research shows that vocal music was dominant, and perhaps the only accepted form of music in early church practice. The earliest recorded Christian Church musical activity was hymn singing (Grout, 2006), the singing was in unison,
and without instrumental accompaniment (Qualben, 2008) because “instrumental music was thought unfit for religious services.”

In addition, analyses of performance practices also give interesting insights into the role of vocal and instrumental music. As already mentioned in the first chapter, music of the gayaki ang in Indian is based on vocal style. Musicians create techniques and try to sing through their instrument like human voice; it is an emulation of a vocal style on instrument. For example the sideway pulling of strings to vary the tension allows sitar players to increase the flexibility of imitating human voice. Erhu, Chinese bowed-string instrument, uses vibrato, glides, and pushing strings to imitate human voice. Pakistani qawwali harmonium artists learn the instrument by singing the melodies first and imitate the melodies through harmonium, and the idea is so the harmonium should like (or at least try to imitate) voice. Another related phenomenon is that flute players from Siberia to South African literally sing through their instruments, thereby producing a special timbre that is both instrumental and vocal at the same time. Instruments in some cultures (e.g. Japanese and Chinese) were made for imitating the nature sounds, for example the birds singing or the sound of waterfall. However, according to the philosophy of these cultures there is no essential separation between humans and nature; natural and human phenomena should be regarded at the same level. These ‘singable’ instruments suggest that human voice was and is very important as an ideal and model in the development of instrumental musics.
In addition, several musical genres have pieces with irregular phrase structures, e.g. composition of Chinese Silk-bamboo ensemble and Chinese solo music. These irregular phrase structures can be understood to indicate that the pieces were originally vocal forms, because vocal singing or speech are frequently characterized by irregular phrase lengths.

Another interesting observation about ‘singing the instrument’ is from Glenn Gould, a pianist who is famous for his interpretation of J.S. Bach’s music. Although Bach’s piano compositions were not vocal music, Gould always sung along when playing. In fact, he was not the only musician who sung while playing. Although Gould claims his singing was unintentional, perhaps the singing opens another possibility to interpret his piano music (Bazzana, 2004) and perhaps there is something in this music that can only be presented by the voice but not by instruments.

Finally, the pitch range of many early melodic instruments in the world is around human voice range, or at least not too far away from this range (e.g. bawu, ancient sitar). This may suggest that the important influences of human voice on instruments develop. In other words, vocal music may have developed earlier, whereas instrumental music, especially those on melodic instruments, was developed later and under the influences of vocal music.
10 Impact and future studies

10.1 Implications for music theory and musicology

If vocal and instrumental musics are processed differently, as indicated by our experiments, then we need to rethink how we actually understand and analyze music. Most music theories do not consider differences between vocal and instrumental music to be of importance, and theorists tend to think they are essentially the same and can be treated in the same way. Especially in Western music vocalists are expected to sing in a way instrumental music is performed; for example, to hold a steady pitch for a note, keep the same timbre for the whole register, or have symmetric meter structure composition for vocal music. These may be realized easily on musical instruments, but are not common in other vocal genres around the world.

The vocal music in Western classical music has been influenced very much by instrumental music. One obvious reason for this is the usage of notation. Notations seem to suggest that a ‘note’ presents a specific pitch, a rationally defined rhythmic value, and a fixed timbre and dynamic. It may be possible for some instrumentalists to follow the notes, but vocalists cannot produce pitches and rhythms as consistently as they are implied by notations.
Most theorists prefer to analyze music through notation, and notations become very critical in the whole analysis. However, as notations have not “arisen as an analytical tool but as a memory and organizational aid within and for the music of a specific culture (Will, 1998),” the analysis of music from notation is very questionable. If instrumental and vocal music are produced and processed very differently, how can they be notated in the same way? “If music theorists really wanted to analyze how music sounds, and only how it sounds, they would put away their scores and forget all about notes (Cook, 1993).” Not only are they not concerned with how music sounds but they are also not concerned with how human process music.

Theorists should not assume all musics could be played, analyzed, and processed in the same way as suggested by notation. In addition, musicologists would profit from recognizing that vocal and instrumental musics have developed from different roots and are subject to different cognitive processing – even though they are now considered just two sides of the same coin, music, represented by one notation system.

10.2 Suggested future research

Participants in the present study were only English-speaking musicians. In no way did we exclude the possibility that culture and language background could influence our results, but, as this was a first study on differences in processing of vocal and instrumental rhythms, we wanted to make sure that we had a group of participants with a homogenous language background (in comparison to the language of the song
examples). In addition, we had only six participants for fMRI scans, we sought to reduce the number of intervening independent variables as much as possible. In the future, if funding allows us to perform more scans and continue the present study, it would certainly be important and interesting to explore if and how culture (e.g. training, professional status) and language experience (e.g. syllable-timed, stress-timed languages) influence rhythm processing.

Another question that arises from the present study, which uses only one percussive instrumental sound, is how the results will change if we use other instruments. Do our results hold only for percussion instruments or for all types of instruments? It would be interesting to expand the experiment to include melodic instruments, for example, to experiment with one of those ‘voice’ like instruments (e.g., sitar and erhu), and see how people process those ‘singable’ instruments in comparison with the human voice.

10.3 New research perspective to ethnomusicology

As described in last chapter, world music suggests that in many cultures vocal music plays a more important role than instrumental music. The experimental results of present study provide an additional explanation for why vocal music seems to be so dominant in musical practices around the world. In addition, this study successfully connected sciences and the humanities and links the formulations of hypotheses with interpretations of musical phenomenon. Through this mutual exchange the cognitive methodology can provide a new way for research in ethnomusicology that complements
the standard approach of fieldwork studies. There is no question about the importance of fieldwork, but, as already outlined earlier, cognitive studies are able uncover aspects of human behavior that remain inaccessible to fieldwork approaches.

There is no doubt that music is a universal activity. Every culture and society has something like ‘music.’ However, a common statement of “music is the universal language of mankind” suggests that the music (any music) can be understood by all mankind. This quote is from the American poet Henry Longfellow. He wrote this sentence after he returned from a trip to Europe. Now this quote is widely used and often meant to indicate that all music can be treated as one and same phenomenon. I do not agree with the statement and do not believe that one can fully understand all music genres of the world in one and the same way. The evidence for differences between vocal and instrumental music, as forwarded in this thesis is an additional factor to caution us against viewing music as as unitary phenomenon that is globally perceived and understood in a uniform manner.

The aim of this dissertation was not to ‘prove’ a new theory on the origin of music. However, the experimental results of this study are in support of the idea of multiple origin of human music. It is hoped that the discourse ensuing from these ideas brings musicians, (ethno-) musicologists, theorists, and other music scholars to acknowledge the necessity and usefulness to distinguish between vocal and instrumental music in our thinking about music as well as in our music making.
Reference


Conference on Chinese linguistics, vol. 1 (NACCL-20, Columbus, OH: The Ohio State University, 2008), p 205-211.


