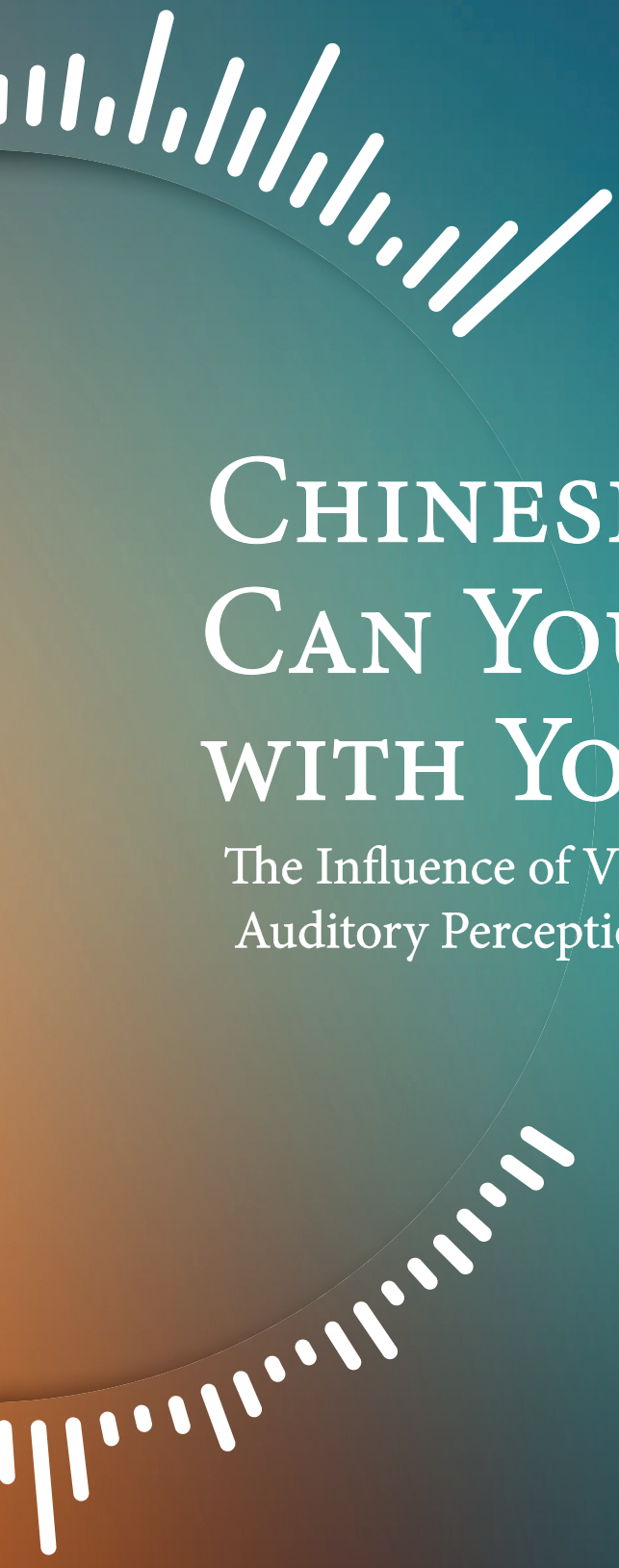
Chinese Tones: Can You Listen With Your Eyes?

Han, Yueqiao

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record



Chinese Tones: Can You Listen with Your Eyes?

The Influence of Visual Information on Auditory Perception of Chinese Tones

Yueqiao Han

Chinese Tones: Can You Listen with Your Eyes?

The Influence of Visual Information on Auditory Perception of Chinese Tones

Yueqiao Han

Financial support was received from China Scholarship Council (CSC).

ISBN: 978-94-6423-261-5
Printed by: ProefschriftMaken | www.proefschriftmaken.nl
Cover design, and layout: Bregje Jaspers | ProefschriftOntwerp.nl
Layout inspired by: http://www.martijnwieling.nl/

Chinese Tones: Can You Listen with Your Eyes?

The Influence of Visual Information on Auditory Perception of Chinese Tones

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University
op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk,
in het openbaar te verdedigen ten overstaan van een door het college voor
promoties aangewezen commissie in de Aula van de Universiteit
op vrijdag 18 juni 2021
om 10.00 uur

door

Yueqiao Han

geboren te Baoding,
China

Promotor:

prof. dr. M.G.J. Swerts (Tilburg University)

Copromotores:

dr. M.B.J. Mos (Tilburg University)
dr. M.B. Goudbeek (Tilburg University)

Leden promotiecommissie:

prof. dr. A. Chen (Utrecht University)
prof. dr. Y. Chen (Leiden University)
prof. dr. D. Burnham (Western Sydney University)
prof. dr. ing. H. Mixdorff (Beuth University Berlin)
prof. dr. J.H.M. Vroomen (Tilburg University)

To *speakers of non-tonal languages*

Acknowledgments

There is never any way to thank all the people whose physical or spiritual presence made this dissertation possible, but I'm going to give it a try anyway.

There are two parties that I would like to acknowledge first: one is my sponsor: the China Scholarship Council (CSC), without whose financial support my journey would be crippled at the beginning. The other one is my supervisors: Marc, Maria, and Martijn. Although it's been said by so many Ph.D. students, and it's kind of a cliché, I still have to repeat after them: I wouldn't have made it to the end without the help and support from them. I don't believe they have co-supervised many students before me, and I'm the lucky one who had to have all three of them, but I would advise them to do this more in the future, because I have experienced it as such a fruitful and powerful combination.

Marc, you always know how to help me with figuring out an interesting introduction and composing an insightful discussion. You are so good at telling a story in a paper by threading through all the scattered information that it's not surprising that you have already published two books. Researcher, writer, and musician (you are the best saxophone player I ever know), how cool even to own one of these titles, let alone all of these!

Maria, your comments are always the first ones to enter my mailbox. You are the one who is never shy to put your fingers on the fatal flaw in my arguments and speak your mind. You are the one who made me sweat when I could not give a satisfactory answer, and you are also the one who motivates me to get better and stronger. You are a master of managing time and work. I learned a lot from you.

And then there is Martijn, the savor. You are the one who joined the team when I was desperate for a hands-on data analysis experience. You are joyful and kind: you always cheer me up when I am worried and you definitely saved my ass from the data-analyzing odyssey. I honestly don't think I would've finished this without your wonderful help!

I would like to extend my thanks to the members of my committee: prof. dr. Aoju Chen; prof. dr. Yiya Chen; prof. dr. Denis Burnham; prof. dr. ing. Hansjörg Mixdorff; and prof. dr. Jean Vroomen. They took their time reading my papers, gave elaborate comments and delivered an assurance for my work. I am very grateful for your work.

There are many people who have helped me along this journey, and I want to single out some of them. Ad Backus, thank you for your comments on my earliest paper and my completed dissertation, and your moral support while I was writing my dissertation. Thiago, thank you for working together with me on my last paper, your expertise made that one happen. Yan Gu, Veronique, Giovana, Mariana, Loes, and Emmelyn, thank you for reading my dissertation and preparing me for the defense; it was very useful and much appreciated. Nadine, Alexandra, and Mariana, thank you for being wonderful friends and helping me with organizing my graduation session.

Another thank you goes out to all my kind colleagues in DCC for treating me well and creating such a wonderful working place: Lauraine, Jacqueline, Alex, Rein, Naomi, David, Tess, Ruben, Debby … My apologies for not naming out everyone. I am grateful that I have made close friends with many of you. In addition, I would like to thank the colleagues I have worked together with for the PhD council and TiPP (Tilburg PhD Platform). Such a pleasant and fruitful experience to have alongside my PhD track!

I am greatly indebted to the participants who took part in my studies and the people who helped me during the data collection. My special thanks goes out to Marlon Titre and his colleagues in Fontys for providing me all the help I needed to collect data for my study.

Even with the professional support of all these people, I would not deliver my dissertation as well as it turned out without the support and encouragement from my families: I would like to take this opportunity to share my deepest gratitude to my husband, Jaap. He has always helped me in all the ways he can. Without his constant proofreading and extraordinary graphing skills, my dissertation would not have been as good as this final version. "And yes, he even proofread these acknowledgments". As well, I am grateful for the presence of my angel, Marin. "Thank you for being such an amazing daughter. You are sweet, smart and growing up so fast. You give me extra strength and determination to finish this dissertation, because being your mom is the most powerful thing I can receive". There is no way that this journey could be completed without the support from my dearest sisters and brother. They have been always supporting me with everything they could. For them, I am always their youngest sister and their love never stopped, even though we are at the different side of the globe.

Most importantly, I have to thank my parents, who silently accepted my decision of leaving my homeland to pursue a study that they never really understand. From the beginning they gave me every opportunity to follow my own path, and I will always be grateful to them. I know my father is proud of me. I also know my mother couldn't be happier with everything I accomplished. Thank you, thank you so much.

Maastricht, April 26, 2021

Julie

Contents

Chapter 1

Chapter 1

General Introduction

This dissertation is a study on the linguistic use of tone. More than half of the languages spoken in the world (60%-70%) are so-called tone languages. Unlike most European languages, which rely primarily on phonological distinctions between consonants and vowels to distinguish word meanings, tone languages, such as Mandarin Chinese, additionally use changes in tone for marking lexical distinctions. Because of its unfamiliarity, tone is known to be difficult to learn for western speakers. This dissertation investigates possible ways to ameliorate the perception of tone for tone-naïve speakers. It sets out to examine the factors that potentially promote efficient perception of Mandarin Chinese tone. To be more specific, this dissertation looks into the contribution of visual information (in particular, potential cues displayed by a speaker's face) to Mandarin Chinese tone perception for tone-naïve perceivers, as well as that of other factors, such as differences in speaking styles of the speaker (natural vs. teaching speaking style), and musicality of the perceivers (musicians vs. non-musicians). These different variables are investigated in a task of Mandarin Chinese tone identification. Moreover, this dissertation also contains a computational study that compares the relative contribution of acoustic information and visual information to tone perception and tone classification. In this chapter, I sketch some background information about tone, especially tone in Mandarin Chinese, embark on the research questions addressed in the dissertation, and give an overview of the studies reported in this thesis, including some considerations of the relevant methodological aspects.

1.1 Tone in Chinese

A language is a "tone language" if the pitch of the word can change the meaning of the word. This means it should not just change its nuances (such as specific emotional or attitudinal connotations of a word), but its core meaning (Yip, 2002). In Mandarin Chinese, for example, when the syllable [ma] is produced with a rising tone, it means "hemp", whereas it means "scold" when produced with a falling tone. In many languages, speech pitch may be modulated for pragmatic purposes: in American English, for instance, using high or low pitch on a word like "Okay" can indicate whether an utterance is intended as a

question or a confirmation. However, such usage is different from the way tone is exploited in a language like Mandarin Chinese, since the core meaning in American English of the word is not changed.

Tone languages consist of nearly 70% of the world's languages, and they are extremely common in Africa (e.g., Yoruba), East and South-East Asia (e.g., Thai), and Central America (e.g., Mixtec) (Yip, 2002). Most of the European languages are not tonal, but there are exceptions like Swedish, Norwegian, Serbo-Croatian, and a few Dutch Limburgian dialects in which tone can also be used to mark lexical contrasts.

Of all those tone languages, Chinese is spoken by the largest population by far (total users in all countries in 2015: 1,107,162,230[1]). Under the general banner of Chinese, eight major language/dialect groups are subsumed: Mandarin, Wu, Yue (Cantonese), Xiang (Hunan), Gan (Jiangxi), Kejia (Hakka), Southern and Northern Min. Although they do share a great deal in common, such as syntax, not one pair of languages is mutually intelligible. The mutual unintelligibility is mostly due to differences in phonology (Bao, 1990, 1999). Mandarin originated in North China and is spoken across most of northern and southwestern China. The Mandarin dialect group is spoken by more people and over a larger geographical area than any other major dialect group (65% of the Chinese population in 2017, estimated by Ethnologue).

In contemporary linguistics, Mandarin tones are often described in terms of pitch height and pitch shape. Accordingly, there are the four main distinctive Mandarin tones[2], conventionally numbered 1 to 4: tone 1: high-level (5-5); tone 2: mid-rising (or mid-high-rising; 3-5); tone 3: low-dipping (also low-falling-rising or mid-falling-rising; 2-1-4); and tone 4: high-falling (5-1) (Chao, 1930)[3].

1.2 Visual information in tone

Tone is an acoustic phenomenon: listeners do not need to see speakers to be able to understand them (e.g., a conversation can take place via the phone). Actually, and interestingly, the more tonal the language, the greater the reliance on auditory information by listeners. Sekiyama and Burnham (2008) explained that as tonal languages (and semi-tonal languages, such as Japanese) having fewer phonemes (consonants, vowels and syllables) and a simpler syllabic and phonological structure compared to English. Because of this, the lip-reading information may be used less in speech/tone processing.

[1] https://www.ethnologue.com/language/cmn.

[2] There is a fifth tone, a neutral tone, which functions on grammatical level and cannot appear on single syllable words.

[3] The numerical substitute has been commonly used for tone contours, with a numerical value assigned to the beginning, end, and sometimes middle of the contour. The numbers 1-5 refer to relative pitch differences; they are not absolute values, as they will vary from speaker to speaker.

Mandarin Chinese, which has an elaborate tonal system, has been shown to have clear acoustic correlates, notably in the form of pitch and pitch contour. In particular, fundamental frequency (F0) patterns (both height and contour) and the direction of pitch, can distinguish the four main distinctive Mandarin tones. Other acoustic variables, such as duration and amplitude, can also be perceptually informative (Chen & Massaro, 2008; Ryant, Yuan, & Liberman, 2014), but to a lesser extent than fundamental frequency. Therefore, acoustic information is essential in (Mandarin) tone perception and accordingly listeners (at least native listeners) rely greatly on it when it is available.

At the same time, the way we perceive speech can be influenced by visual factors: it is a multisensory/multimodal process. What we *hear* can be affected by what we *see* (Campbell, Dodd, & Burnham, 1998; Han, Goudbeek, Mos, & Swerts, 2018, 2019; Rosenblum, 2008). For instance, seeing the face of the speaker normally helps the listener perceive speech better (Bailly, Perrier, & Vatikiotis-Bateson, 2012; Hirata & Kelly, 2010; Sumby & Pollack, 1954), especially in noisy environments (e.g., Burnham, Lau, Tam, & Schoknecht, 2001; Mixdorff, Hu, & Burnham, 2005b). Similarly, seeing the face of a speaker also aids hearing impaired listeners decoding the auditory speech signal (Desai, Stickney, & Zeng, 2008; Smith & Burnham, 2012).

Accordingly, it stands to reason that tone perception is also more than a purely auditory event. The current dissertation focuses on visual information displayed by motions/cues from lips, face, head, and neck. Whether or not there is visual information in tone has been answered by previous studies (e.g., Burnham, Ciocca, Lauw, Lau, & Stokes, 2000; Burnham et al., 2001; Mixdorff & Charnvivit, 2004; Mixdorff, Charnvivit, & Burnham, 2005a; Mixdorff et al., 2005b) that confirm that lexical tones are marked by visual information as well. Physiological studies (e.g., Xu & Sun, 2002) suggest certain restrictions with respect to the coordination of the laryngeal and articulatory systems, which may lead to visual cues for tones (Mixdorff et al., 2005a). Because our mouth, face, and head need to move in a certain way to produce a given tone, the amplitude (range) and the length (duration) of the visible articulations change. For example, in Mandarin Chinese tones, there are clear differences in the duration of the vowels and the amplitude across tones: tone 3 usually has the longest vowel duration, while tone 4 tends to be the shortest; the amplitude for tone 3 is usually the lowest one, whereas tone 4 normally has the highest amplitude (Tseng, 1981). It makes a lot of sense that such articulatory changes, for instance the amplitude (loudness) and the length of the articulation, are visually displayed in movements in and by the face (Kim & Davis, 2001; Reid et al., 2015). When speakers want to convey information about tone (the pitch contour, for instance), facial cues (along with gestures) may thus represent a useful visual resource they resort to alongside the acoustic information,

consciously or unconsciously, to produce the different melodic configurations (Rosenblum, 2008; Swerts & Krahmer, 2008; Zheng, Hirata, & Kelly, 2018).

A number of studies have explored the nature and locus of the visual cues in tone production and perception, and have revealed fairly reliable configurations of visual cues related to tone acquisition, even though exact visual cues have not been clearly defined yet. For instance, strong correlations between head movements and F0 were observed by Yehia, Kuratate, and Vatikiotis-Bateson (2002). Similar visual cues that relate to more general movements of the head and/or eyebrows have previously been reported to function as correlates of larger-scale prosodic structures in other languages, for example, quick movements of the head that co-occur with pitch accents (Burnham et al., 2006; Krahmer & Swerts, 2007; Vatikiotis-Bateson & Yehia., 1996).

Although there is visual information when speakers produce tones, whether and how this information is picked up by perceivers has been attracting scholars' attention for the past two decades. Burnham et al., (2000) tested native identification of (six) Cantonese tones with auditory- only, visual-only, and auditory-visual modes, this being the first empirical study on the cue value of visual information for lexical tone. Their participants' performance of tone perception in the visual only condition is significantly above chance level, which provides evidence that there is indeed visual speech information for lexical tone perception. Since then, more studies on visual and audio-visual tone perception have been conducted. In 2001, Burnham et al., conducted a same-different discrimination study on Cantonese tones, in which native Thai and Australian English speakers also performed significantly better than chance under visual-only conditions. Also, Chen and Massaro (2008) found that the performance of native Mandarin speakers in visual lexical-tone identification was statistically significant. Visual facilitation for tone identification has been especially found for speech in noise for both Mandarin (Mixdorff et al., 2005b) and Thai (Mixdorff et al., 2005a; Burnham et al., 2015).

Chen and Massaro (2008) found that visual tone identification improved significantly after the participants were taught to pay attention to the visual movements of the neck, head, and mouth. Specifically, Mandarin Chinese tone identification has been found to mainly depend on the (intensity of the) movements of the mouth, head/chin, and neck. The amount of visual information involved in individual Mandarin Chinese tones varies between tones: there is little to no activity for tone 1, some activity for tone 2 and tone 4 (though very brief for tone 4), with tone 3 having the most activity, namely a dipping head/chin. Duration (time) differences between the tones may be caused by variation in the movements of the mouth, as more complex movements would require more time to be realized. More recent audio-visual research on tone languages from Burnham et al., in 2019 added larynx motion (in addition to head motion) as a possible cue for Thai tone classification, and

they found positive evidence that this type of motion is important for Thai tone production. Further research will be vital to describe visual tone cues more precisely, in both perception and production (Reid et al., 2015).

In general, visual speech information is known to benefit speech perception. For instance, an early study conducted by Sumby and Pollack in 1954 showed that seeing the speakers' face helps the listeners' intelligibility. More specifically, for the perception of tone, the visual facilitation mainly appears under difficult listening conditions (e.g., impaired listeners or noise-masked auditory signal) (see Campbell, et al., 1998, and Bailly et al., 2012, for a comprehensive collection of studies). As for the extent to which auditory-visual information facilitates or improves tone identification compared to auditory-only information (i.e., the superiority of bimodal performance compared to unimodal performance), it differs widely across individuals' experience (Burnham et al., 2015; Grant & Seitz, 1998). Furthermore, the benefits of visual/facial information for tone perception depend strongly on context, and in particular on the availability of a clear and reliable acoustic signal. In situations where such a signal is available, extra visual information may actually distract the perceivers instead of facilitating their tone perception, since they are reluctant to use the visual information when acoustic sources are available and reliable. For example, Burnham et al. (2001) have found that in an experiment using clean speech, Australian English speakers performed better in a task of identifying Cantonese words that differed only in tone in the auditory-only (AO) condition than in the auditory-visual (AV) condition (where they also had access to lip and face movements).

Similar results also appeared in another study concerning visual cues in tone perception conducted by Mixdorff, Hu, and Burnham (2005b). In their study, native Mandarin speakers identified Mandarin tones in various auditory and/or visual conditions (clean, reduced, and masked audio-only/audio-visual). They found that adding visual information in the clear and devoiced auditory conditions was not particularly helpful. However, tone perception improved significantly in the babble-noise masked condition. The authors speculated that the absence of a facilitating effect for visual information on tone identification may be due to a ceiling effect for native speakers in clear audio conditions: auditory information suffices for quick and correct identification of tones, unless this information is compromised, that is, under low speech-to-noise ratios, in which case visual information is beneficial. Smith and Burnham (2012) found that tone-naïve listeners outperformed native listeners in the visual-only condition in a task of Mandarin tone discrimination, additionally suggesting that visual information for tone may be underused by normal-hearing tone language perceivers.

Overall, the beneficial effect from visual information appears to be more obvious on the segmental level (i.e., consonants and vowels) than on the

suprasegmental level (i.e., tone), and more obvious on the non-native perceivers than on the native perceivers.

1.3 Elements and variables in current studies

While the way tones are acquired by listeners has attracted some scholarly attention (e.g., Burnham et al., 2000; 2001; Francis, Ciocca, Ma, & Fenn, 2008; Hao, 2012; So & Best, 2010), detailed knowledge of the factors that promote efficient acquisition is lacking. The current studies investigate several factors that are potentially important for the acquisition of tones, but have not yet been studied in a systematic way, or have not been combined in an integrated approach. These factors can be categorized into two groups: (1) contextual factors, such as the auditory, visual or audio-visual modality in which speech is presented, and speaking style of a speaker who is producing speech in a natural or teaching manner; and (2) individual characteristics, related to differences between tone-native and tone-naïve perceivers, and to perceivers with and without musical backgrounds.

1.3.1 Contextual factors

As argued in the preceding section, previous studies (e.g., Chen & Massaro, 2008) have shown that visual information does play a role in Chinese tone perception, and that the different tones correlate with variable movements of the face. To extend the existing body of literature, Chapter 2 in this dissertation looks deeper into the effect of visual cues in a speaker's face on tone identification. The hypothesis is that learners who can see the speakers (audio-visual condition) outperform those who only have access to auditory information (audio-only condition). Since perceivers (at least, tone-naïve perceivers) seem to rely both on auditory and visual information in speech perception (Bailly et al., 2012; Burnham et al., 2001; Calvert, Spence, & Stein, 2004; Campbell et al., 1998; Massaro, 1998), the relative strength of these two factors and possible interactions between them was exploited as well. The impact of visual cues on the segmental level has been demonstrated with the classic McGurk effect (McGurk & MacDonald, 1976): observers perceived an auditory [ba] paired with a visual [ga] as "da" or "tha". This shows that auditory speech perception changes with simultaneously presented incongruent/ discrepant visual information of the speaker's face. In other words, access to visual information about the source of speech can have clear effects on speech perception, as it alters the perception of speech. One of the goals of this dissertation is to investigate a possible existence of auditory-visual confusion (i.e., McGurk effect) beyond the segmental level, that is, the four Mandarin Chinese tones (suprasegmental level) (Chapter 4). More specifically, the

relative contribution of auditory and visual information was compared during Mandarin Chinese tone perception with congruent and incongruent auditory and visual materials for speakers of Mandarin Chinese and speakers of non-tonal languages. We further explore the contribution of visual cues by adding them to a computational model for tone classification that has so far been based on conventional acoustic features only (Chapter 5). By comparing automatic and human classification of Mandarin Chinese tones, the representativeness of our models as models of tone learning is assessed.

The second contextual factor concerns possible adjustments speakers of the native language make when they talk to learners of their language. There is evidence that speakers adapt their speaking style to their audience and to the communicative context. A well-known example of this is infant-directed speech (IDS), where adults adapt their speaking style in the presence of the children (Burnham, Kitamura, & Vollmer-Conna, 2002; Fernald & Kuhl, 1987; Kuhl et al., 1997). IDS has been hypothesized to aid the learning process (Kuhl et al., 1997; Thiessen, Hill, & Saffran, 2005). Similarly, a native speaker who is addressing a non-native listener may adapt their speech to improve learning and understanding. In a teaching setting, they may, for example, be more inclined to speak slowly and in a more hyperarticulated manner (Bradlow & Bent, 2002; Smiljanić & Bradlow, 2007, 2009). Assuming that a teaching style that attends to the needs of learners may also make tonal contrasts more salient, my dissertation also aims to study whether a hyperarticulated speaking style helps learners to perceive tonal information (Chapter 2, 3 and 5).

1.3.2 Individual differences between perceivers

Visual information is mainly relevant for native speakers, and, in general, speakers of tone languages focus more on auditory information than speakers of non-tone languages. This then raises questions about its use by and usefulness for tone-naïve people, who (have to) acquire the tones. Therefore, I chose to mainly focus on tone-naïve participants in our tone identification experiment in order to establish the tone acquisition process (Chapter 2, 3, 4 and 5), and to avoid the emergence of ceiling effects on tone identification that would typically appear from native speakers. Most importantly, my studies on tone-naïve speakers can contribute to the field of tone learning for second/foreign language speakers.

Furthermore, the unfamiliarity with tone in many Western speakers makes tone languages ideally suited to examine the influence of musical experience on language acquisition (Marie, Delogu, Lampis, Belardinelli, & Besson, 2011), given the fact that both the perception of native (Schön, Magne, & Besson, 2004) and foreign language speech (Marques, Moreno, Luís Castro, & Besson, 2007) have been reported to benefit from musical experience (Marie et al., 2011;

Milovanov, Huotilainen, Välimäki, Esquef, & Tervaniemi, 2008; Milovanov, Pietilä, Tervaniemi, & Esquef, 2010). This dissertation aims to explore whether or not musical expertise also helps tone-naïve listeners to correctly identify Mandarin Chinese tones (Chapter 3). Because of extensive musical training, musicians are particularly sensitive to the acoustic structure of sounds (i.e., frequency, duration, intensity, and timbre parameters). This sensitivity has been shown to influence their perception of pitch contours in spoken language (Schön et al., 2004), but the extent to which musicians are affected by the presence of (exaggerated) visual information during speech perception has remained largely unexplored. Given their extensive training to analyze the acoustic signal, they might not be as inclined to use visual cues as non-musicians and therefore might benefit less from the added visual information. We hypothesize that musicians may still benefit from the added visual information for the Mandarin tone identification, but that this contribution is likely *smaller* than that for non-musicians.

1.4 Research questions

The aim of this dissertation is to study the value of visual information (over and above acoustic information) in Mandarin tone perception for tone-naïve perceivers, in combination with other contextual and individual factors. Moreover, this dissertation exploits the relative strength of acoustic and visual information in tone perception and tone classification. The next four chapters present four studies aiming to answer these research questions. Generally, Chapter 2 and 3 report on empirical studies setting out to investigate to what extent tone-naïve perceivers are able to identify tones in isolated words, and whether or not they can benefit from (seeing) the speakers' face and hyperarticulating speaking style, and their own musical experience. Chapter 4 deals with whether or not there is an audio-visual integration at the tone level in native speakers of Mandarin Chinese and tone-naïve perceivers (i.e., we explored perceptual fusion between auditory and visual information). Chapter 5 studies the acoustic and visual features of the tones produced by native speakers of Mandarin Chinese. Computational models based on acoustic features, visual features and acoustic-visual features are constructed to automatically classify Mandarin tones. More detailed research questions are presented in each empirical chapter.

1.5 Methodology

In order to answer those research questions, there are several methodological aspects that I will briefly introduce. First, both production and perception

studies were included in each empirical chapter of this dissertation. Although the produced tones by native Mandarin Chinese speakers was mainly used to gather the experimental stimuli for the perception test, and perception is the focus of this thesis, the assumption here is that by studying both production and perception we can look into what the perceivers pick up from what the speaker produces acoustically and visually. Native Mandarin Chinese speakers were instructed to produce individual words, while the participants are asked to identify which tone they think they have heard and/or seen (Chapter 2, 3, and 4). The produced stimuli can give us information about what a speaker does (the acoustic information they convey, the visual cues they employ), and the perception results tell us to what extent this information is relevant for the perceivers (Chapter 4 and Chapter 5). An additional reason to look into both production and perception in our study is that it is known from previous work that there is not necessarily a direct relation between speech production and perception (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Casserly & Pisoni, 2010; Baese-Berk & Samuel, 2016). For instance, some acoustic variation, while a systematic and potentially good classifier, may not work well in perception, because it is below a perceptual threshold. Similarly, there is also no direct relationship between tone production and perception (Wang, Spence, Jongman, & Sereno, 1999; Wang, Jongman, & Sereno, 2003).

Second, we recruited (Mandarin Chinese) tone-native and tone-naïve participants with various backgrounds (mainly Dutch). Data from the former group were needed to set a baseline for the perception experiment, and to compare performance in employing visual cues (Chapter 2 and Chapter 4). Questionnaires were used to select and group the participants according to their language background (tone-native and tone-naïve) and musical behavior (musician and non-musician). These questionnaires are commonly used in research on language learning (Chapter 2 and Chapter 3) and the reliability and validity of them have been established. Third, to assess the importance of acoustic and visual features in the process of tone classification and tone perception, Chapter 5 made use of Machine Learning (ML), a subarea of Artificial Intelligence which aims to *learn* how to categorize data by using patterns and inference instead of explicit instructions. By comparing automatic and human classification for Mandarin Chinese tones, the representativeness of our models as models of tone learning can be established.

The data obtained in the production experiments were analyzed with popular software: Praat and The Computer Expression Recognition Toolbox (CERT). Praat 6.0.33 (Boersma & Weenink, 2017) was used to measure the acoustic features. CERT was employed to automatically code facial expressions in the videos. Based on previous literature and the theoretical significance of the features, a set of acoustic parameters and relevant action units (AU) were selected. The analysis of these features provided statistical evidence for

the validity of the experimental stimuli on the one hand, and are the basis of classifying tones on the other hand.

A recurring methodology used in analysing the data obtained by the perception experiments is repeated measurements ANOVAs (in Chapter 2 and Chapter 3), which is widely used in psycholinguistics and serves the goals of the experiments sufficiently (Chapter 2 and Chapter 3). The results were also analysed with mixed-effect models in the R software program (Baayen, Davidson, & Bates, 2008) when more variables needed to be included in the analysis (Chapter 4). Since the last study (presented in Chapter 5) was concerned with a classification problem, Logistic Regression was used to classify the produced and perceived tones. In addition, we also heavily relied on analyses of confusion matrices to gain insight into patterns in how perceivers had categorized the various tones.

1.6 Overview

Having introduced the research questions and some of the methodological aspects of this dissertation, I can now present an overview of the remaining chapters. Chapter 2 to 5 are self-contained texts (i.e., they all have their own abstract, introduction and discussion section), and are based on articles either published in (chapter 2, 3, 4) or submitted (chapter 5) to peer-reviewed journals. Therefore, some overlapping texts between individual chapters and between those chapters and this introduction are unavoidable. The author of this thesis was the main researcher in all studies presented here.

The study presented in **Chapter 2** investigates the effect of visual cues (comparing audio-only with audio-visual presentations) and speaking style (comparing a natural speaking style with a teaching speaking style) on the perception of Mandarin tones by non-native listeners, looking both at the relative strength of these two factors and their possible interactions. Native speakers of a non-tonal language were asked to distinguish Mandarin Chinese tones on the basis of audio (-only) or video (audio-visual) materials. In order to include variations, the experimental stimuli were recorded using four different speakers. Participants' responses and reaction times were recorded. The proportion of correct responses and average reaction times were reported.

Continuing the exploration of the potential factors in tone perception, **Chapter 3** is concerned with the effects of musicianship of the participants (combined with those of modality) on Mandarin tone perception. A considerable number of studies have shown that musical ability has a positive effect on language processing. Extending this body of work, this study investigates the effects of musicality and modality on Mandarin tone identification in tone-naïve participants. To examine the effects of visual information in speech, Mandarin tones were presented in auditory-only or auditory-visual modalities

to participants with or without musical experience. The Goldsmiths Musical Sophistication Index (Müllensiefen, Gingras, Musil, & Stewart, 2014) was used to measure the musical sophistication of each participant. A linear regression analysis was conducted to find out whether a specific musical ability/skill as measured by the subscales of the Gold-MSI is related to successful tone identification. Since the effects of the two independent variables might vary among tones, the effects for each tone were subsequently assessed individually in the study.

Chapter 4 focuses on comparing the relative contribution of auditory and visual information during Mandarin Chinese tone perception. Two questions were investigated in this chapter: the first question is whether a McGurk effect can also be discerned at the tone level in native speakers of Mandarin Chinese. Secondly, how visual information affects tone perception for native speakers and non-native (tone-naïve) speakers. To answer these questions, various tone combinations of congruent (A_xV_x) and incongruent (A_xV_y) auditory-visual materials (10 syllables with 16 tone combinations each) were constructed and presented to native speakers of Mandarin Chinese and speakers of non-tonal languages. Accuracy, defined as the percentage correct identification of a tone based on its auditory realization, was used as the dependent variable. In general, there are two assumptions: one is that (native and tone-naïve) participants mainly depend on auditory information when they have to identify Mandarin Chinese tones. Both groups of participants therefore are expected to identify the congruent stimuli more accurately than the incongruent ones. The other one is that (congruent) visual information would facilitate speech perception, especially for perceivers who lack comprehensive knowledge of the language (tone-naïve participants), while this additional value of visual cues would be less important for native participants. Furthermore, when participants are presented with the incongruent experimental materials, there are three types of possible outcomes of how the cues from different modalities are combined: non-integration, integration, and attenuation.

Chapter 5 is another chapter to zoom in on the relative importance of auditory and visual information for tone-naïve perceivers, but from the aspect of tone classification. Moreover, this study examines what the perceivers pick up (perception) from what the speaker does (auditory signal, facial features) by studying both production and perception. To be more specific, this chapter sets out to answer: (1) which acoustic and visual features of tones produced by native speakers could be used to automatically classify Mandarin tones. Furthermore, (2) whether the features used in tone production are similar to or different from the ones that have cue value for tone-naïve perceivers when they categorize tones; and (3) whether and how visual information (i.e., facial expression and facial pose) contributes to the classification of Mandarin tones over and above the information provided in the acoustic signal. To address

these questions, four Mandarin speakers were videotaped while they produced ten syllables with four Mandarin tones (i.e. 40 words in two styles - natural and teaching), totaling 160 stimuli (the same stimuli in Chapter 2). These audio-visual stimuli were subsequently presented to 43 tone-naïve participants in a tone identification task (the same data from non-musicians in Chapter 3). Basic acoustic and visual features were extracted. We used various machine learning techniques to identify the most important acoustic and visual features for classifying the tones. The classifiers were trained on produced tone classification (given a set of auditory and visual features, predict the produced tone) and on perceived/responded tone classification (given a set of features, predict the corresponding tone as identified by the participant).

Finally, **Chapter 6** provides a general discussion of the main findings, and specifies implications for future research.

Chapter 2

Chapter 2

Effects of Modality and Speaking Style on Mandarin Tone Identification by Tone-naïve Listeners

Abstract. Although the way tones are acquired by second or foreign language learners has attracted some scholarly attention, detailed knowledge of the factors that promote efficient learning is lacking. In this article, we look at the effect of visual cues (comparing audio-only with audio-visual presentations) and speaking style (comparing a natural speaking style with a teaching speaking style) on the perception of Mandarin tones by non-native listeners, looking both at the relative strength of these two factors and their possible interactions. Both the accuracy and reaction time of the listeners were measured in a task of tone identification. Results showed that participants in the audio-visual condition distinguished tones more accurately than participants in the audio-only condition. Interestingly, this varied as a function of speaking style, but only for stimuli from specific speakers. Additionally, some tones (notably tone 3) were recognized more quickly and accurately than others.*

2.1 Introduction

For many second language learners, the ultimate goal of learning a language is to be able to communicate like a native speaker. Acquiring a new language entails that a whole gamut of linguistic structures needs to be learned, including grammatical, lexical and phonological characteristics, as well as pragmatic aspects of language use. This paper focuses on acquiring

*This chapter is based on: Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2018). Effects of modality and speaking style on mandarin tone identification by non-native listeners. *Phonetica*, 76(4), 263-286. https://doi.org/10.1159/000489174

specific phonological properties of a language, namely tones in Mandarin Chinese. Chinese tones serve to distinguish word meanings. For instance, if the Mandarin Chinese syllable /ma/ is produced with a rising tone, it means "hemp", whereas it means "scold" when produced with a falling tone. Obviously, tonal information represents a crucial aspect of this language's sound structure, and needs to be learned by a second language learner.

While the way tones are acquired by listeners has attracted some scholarly attention (Burnham et al., 2000, 2001; Francis et al., 2008; Hao, 2012; So & Best, 2010), detailed knowledge of the factors that promote efficient learning is lacking. In the current study, we investigate two factors that are potentially important for the acquisition of tones, but have not been studied yet in a systematic way, and have not been combined in an integrated approach.

First, we explore the effect of visual cues in a speaker's face on tone identification by tone-naïve listeners. In most of our daily interactions, we both hear *and* see our conversation partners: whenever visual information is available, observers use it to decode what they hear (Davis & Kim, 2004; Hazan et al., 2006; Navarra & Soto-Faraco, 2007). However, while it has been shown that speech perception in general is affected by such visual information, the added value of facial expressions for tone perception is context-dependent. It has even been argued that the extra visual information from the face may actually distract the listeners from accurate tone perception, since listeners are reluctant to use the visual information when acoustic sources are available and reliable (Burnham et al., 2001). To gain an insight into the possible added value of visual information and identify under what circumstances listeners use visual information, we test whether learners who can see the speakers outperform those who only have access to auditory information (see section 2.1.1).

The second factor concerns the effect of speaking style. Listeners usually encounter tones in two different speaking styles: a natural style, representing the way native speakers speak in most of their daily interactions; and a teaching style, that is, the hyperarticulated manner in which teachers/native speakers address non-native speakers in a teaching context. Assuming that a teaching style that attends to the needs of learners may also make tonal contrasts more salient, the second goal of our study is to study whether a teaching style helps learners to perceive tonal information (see section 2.1.2).

In sum, we look at the effects of visual cues and speaking style on the perception of Mandarin tones by non-native listeners, and we are specifically interested in the relative strength of these two factors and possible interactions between them. In order to examine this, we assess how well listeners who do not have a tonal system in their first language are able to learn Mandarin Chinese tones and how this learning depends on the above-mentioned factors. In the following sections, we discuss earlier work on these two factors, as well as potential variation and differences between speakers, listeners, and tones.

2.1.1 The effect of modality on tone perception

Generally speaking, speech perception is multimodal, which means that it involves information from more than just the auditory modality. Whenever visual information is available, observers use it to decode what they hear (Bailly et al., 2012; Burnham et al., 2001; Calvert et al., 2004; Campbell et al., 1998; Massaro, 1998). Visual information is provided by movements of the lips, face, head and neck. The impact of such cues has been demonstrated with the classical McGurk effect (McGurk & MacDonald, 1976): observers perceived an auditory [ba] paired with a visual [ga] as "da" or "tha". This shows that auditory speech perception changes with simultaneously presented incongruent visual information of the speaker's face. In other words, access to visual information about the source of speech can have clear effects on speech perception, as it alters the perception of speech.

Various studies have already investigated the impact of visual information on speech perception by linking facial cues and gestures (head and/or hand) to speech comprehension. The results have demonstrated the supportive role of visual information for speech perception in face-to-face interaction (Hirata & Kelly, 2010; Sueyoshi & Hardison, 2005; Yehia, Kuratate, & Vatikiotis-Bateson, 2002). One early study, conducted by Sumby and Pollack in 1954, examined the contribution of visual factors to oral intelligibility by manipulating the presence or absence of a supplementary visual display of a speaker's facial and labial movements. Subjects were instructed to select the words they heard (or they thought they had heard) from a furnished list. When the speakers could both be seen and heard, the speech was considered to be more intelligible, in particular when the speech-to-noise ratio was low (i.e., in noisy contexts) or the number of alternatives listeners had to choose from was limited. These results suggest that supplementary visual observation of the speaker improves the intelligibility of oral speech in specific situations. In general, congruent visual information during articulation facilitates speech perception (Cutler & Chen, 1997; Hallé, Chang, & Best, 2004; Ye & Connine, 1999).

More specifically, the role of visual information in tone perception has been the focus of a number of studies (Burnham et al., 2000, 2001, 2006; Chen & Massaro, 2008; Mixdorff et al., 2005a, 2005b; Mixdorff, Lirong, Nguyen, & Burnham, 2006; Mixdorff & Charnvivit, 2004; Reid et al., 2015). These studies have examined the integration of audio and visual information during tone perception and/or production by presenting the experimental stimuli in bimodal (audio-visual) or unimodal (audio-only, visual-only) sensory conditions and in clear or degraded auditory settings (i.e., with smaller or larger speech-to-noise ratios). In general, these studies show that visual information is relevant for lexical tone perception, although the amount of audio-visual

benefit achieved (i.e., the superiority of bimodal performance in relation to unimodal performance) differs widely across individuals (Grant & Seitz, 1998).

Burnham et al. (2000), for example, investigated the perception of Cantonese tones with Cantonese native speakers, who were either phonetically-trained or phonetically-naïve. They found no difference in performance between audio-only and audio-visual conditions and while listeners performed worse in the visual-only condition, they still performed above chance. Interestingly, phonetically-naïve listeners outperformed phonetically-trained listeners, which Burnham et al. attribute to attentional and learning processes. Another study concerning visual cues in tone perception was conducted by Mixdorff et al. (2005b). In their study, native Mandarin speakers identified Mandarin tones in various auditory and/or visual conditions (clear, reduced, and masked audio-only/audio-visual). They found that, in the clear and devoiced auditory conditions, adding visual information was not particularly helpful (similar to the findings of Burnham et al., 2000). However, tone perception was significantly improved in the babble-noise masked condition. The absence of a facilitating effect for visual information on tone identification may be due to a ceiling effect for native speakers in clear audio conditions: auditory information suffices for quick and correct identification of tones, unless this information is compromised, that is, under low speech-to-noise ratios, in which visual information is beneficial.

While all of the studies above have used native listeners as participants, some studies have included non-native participants. For example, Burnham et al. (2001) compared tonal and non-tonal native speakers (Thai and Australian English) in their ability to discriminate Cantonese tones. They found that both groups performed significantly above chance, even in visual-only conditions, confirming that there is visual information in the face for tone discrimination. However, they did not find an advantage among Australian English speakers for audio-visual stimuli in neither clear nor noisy auditory conditions. Thai native speakers, however, did benefit from audio-visual presentation in noisy conditions. In this study, the researchers manipulated the speech-to-noise ratio by adding a certain level of background noise. They concluded that visual cues were more salient for the Thai listeners in degraded auditory settings. These findings suggest that perceivers that have difficulty accessing the auditory material because of noise, hearing impairment or because it is not in their native language, might benefit the most from the supplementary visual information when listening to tones.

To examine the effects of visual speech on tone perception among hearing-impaired perceivers, Smith and Burnham (2012) carried out a Mandarin tone perception task among cochlear implant users. They asked native speakers of Mandarin and Australian English speakers to discriminate between minimal pairs of Mandarin tones in five conditions: audio-only, audio-visual, cochlear

implant-simulated audio-only, cochlear implant-simulated audio-visual and visual-only. They found that, in the visual-only condition, both Mandarin and Australian English speakers discriminated tones above chance levels. As in Burnham et al. (2000), tone-naïve listeners (Australian English speakers) outperformed native speakers of Mandarin Chinese. Their explanation was that the visual information may in fact be underused by native speakers that have come to rely on their auditory abilities for their native language.

Given the mixed effects with respect to the contribution of visual cues to tonal perception and the possible role of ceiling effects, the current study investigates participants who were naïve with respect to tone identification: native speakers of a non-tonal language. This strongly reduces the possibility of ceiling effects when comparing audio-visual and audio-only conditions. We focused primarily on the added value of visual information for tone-naïve listeners in two clear, yet distinct auditory conditions: when speakers employ a "teaching style" specifically geared to non-native listeners or a more natural speaking style, geared towards fellow native speakers. This distinction is discussed in the following section.

2.1.2 The effect of speaking style on tone perception

Adult speakers possess the ability to intuitively and automatically adjust their speaking style to meet the demands of the target audience or the communicative situation (Junqua, 1993; Kuhl et al., 1997; Skowronski & Harris, 2006). They show sensitivity to characteristics of the audience they are addressing (Burnham et al., 2002). To make themselves more intelligible to the listeners, speakers usually articulate in a more "exaggerated" manner: they maximize phonetic contrast, attempt to speak more slowly, more loudly, and more clearly (Smiljanić & Bradlow, 2009). These modifications in speaking style have been discussed extensively as "clear speech" (Ferguson & Kewley-Port, 2007; Smiljanić & Bradlow, 2007, 2009; Uchanski, 2005).

Clear speech modification aims at providing more salient acoustic cues in the speech signal for the listeners to enhance their access to and comprehension of the message (Smiljanić & Bradlow, 2009). Various listeners' populations have been shown to benefit from clear speech, including adults with normal or impaired hearing (Uchanski, Choi, Braida, Reed, & Durlach, 1996; Ferguson & Kewley-Port, 2002; Krause & Braida, 2002; Ferguson, 2004; Liu, Del Rio, Bradlow, & Zeng, 2004; Smiljanić & Bradlow, 2005; Maniwa, Jongman, & Wade, 2008), elderly adults (Schum, 1996; Helfer, 1998), as well as non-native listeners (Bradlow & Bent, 2002; Bradlow & Alexander, 2007; Smiljanić & Bradlow, 2007). The effects of clear or hyperarticulated speech on the perception of tones has been studied in, for example, Infant Directed Speech (IDS). Xu and Burnham (2010) examined hyperarticulated tones in Cantonese IDS and

concluded that tone fidelity is not affected by the exaggerated intonation of IDS. In contrast to the claims that IDS helps in highlighting important aspects of speech (Thiessen et al., 2005), Benders (2013) argues that a hyperarticulated speaking style in IDS might not be helpful to facilitate language learning, but is primarily meant to promote affection between mothers and infants. Similarly, there is substantial literature criticising the didactic notion of IDS by Alex Cristia and her colleagues (e.g., Cristia, 2013; Cristia & Seidl, 2014; Martin et al., 2015).

In the field of second language learning, a hyperarticulated speaking style is commonly associated with "teacher talk" (or "foreigner talk") that teachers use when addressing second language learners in the classroom, anticipating learners' needs for assistance in their attempts at comprehension (Ferguson, 1975, 1981). For instance, the paper of Uther, Knoll and Burnham (2007) concerned a comparison of "foreigner-directed-speech" (FDS), IDS and regular adult-directed speech. The results suggest that linguistic modifications found in both infant- and foreigner-directed speech are didactically oriented, and that linguistic modifications are independent of vocal pitch and affective valence.

With respect to vocal aspects of different speaking styles, more attention has been paid to segmental correlates like vowels and transitions (Llisterri, 1992) than to lexical tone information (Chen & Massaro, 2008). To the best of our knowledge, there are no previous studies that have explored to what extent the acquisition of tones by non-native listeners is affected by the speaking style to which they are exposed, which is all the more surprising given the ubiquitous use of clear speech by foreign language teachers. So, the second goal of our study is to investigate whether exposure to a hyperarticulated speaking style (teaching style) leads to better tone recognition than exposure to normal speaking style.

2.1.3 Variation between speakers and between tones

The primary goals of our study are to investigate the impact of the visual modality and of teaching style on tone identification by tone-naïve listeners. Both factors exhibit inherent variation that will be explored as well. First, there is variation between speakers. Some speakers differ in the clarity of their articulation and some are easier to understand than others (e.g., Gagné, Masterson, Munhall, Bilida, & Querengesser, 1994). In an early study, Cox, Alexander, and Gilmore (1987) recorded three male and three female speakers to test auditory speech intelligibility among groups of normal hearing subjects and found significant differences in speech intelligibility across speakers. In a corpus of conversational and clear speech from 41 speakers, Ferguson (2004) found that female speakers tended to produce more intelligible clear speech compared to male speakers. Similarly, speakers may also vary in the clarity

of the visual cues that they provide (Grant & Braida, 1991; Lesner & Kricos, 1981). For example, in a study presenting visual-only stimuli from six female speakers to a group of normal-hearing subjects, three speakers were judged to be difficult to speechread and three were judged to be fairly easy to speechread (Kricos & Lesner, 1982).

With respect to tone perception, there has been relatively little research on speaker variation (but see Creel, Aslin, & Tanenhaus, 2008; Gagné et al., 1994; Nygaard, Sommers, & Pisoni, 1995). In most cases, only one speaker was employed to produce all experimental stimuli (such as in Burnham et al., 2001, 2006; Mixdorff & Charnvivit, 2004; Mixdorff et al., 2005a, 2005b; Reid et al., 2015). The study of Smith and Burnham (2012), mentioned previously, recruited two adult native speakers of Mandarin (one male and one female), but their study, given the low number of speakers, does not address a possible speaker effect. Chen and Massaro (2008) studied the role of visual information and tone perception and showed that female speakers were easier to understand. In their study, four Chinese native speakers, two male and two female, produced the experimental materials. Mandarin participants identified the tones before and after a learning phase. The results revealed that performance was generally better if the speaker was female. Their explanation for this finding was that female speakers tended to have more salient head/chin movement than male speakers. Based on the above, individual speaker differences should be examined, independently of modality and speaking style.

In addition to variation between speakers, differences between tones should also be taken into account. We aim to see whether the way the tones are acoustically realized is also visually signaled. For instance, vowel duration tends to be the longest for tone 3 and shortest for tone 4; amplitude tends to be lowest for tone 3 and highest for tone 4 (Tseng, 1981; Tseng, Massaro, & Cohen, 1986). These acoustic differences may have visual correlates, for instance in the amplitude and the length of the visible articulations. Physiological studies (Xu & Sun, 2002) suggest certain restrictions with respect to the coordination of the laryngeal and articulatory systems, and these might be responsible for the visual cues for tones (Mixdorff et al., 2005a). Moreover, as far as prosodic features are concerned, it has been shown that there is a strong correlation between head motion and fundamental frequency (Yehia et al., 2002). This suggests that head motion can be used to estimate the fundamental frequency during the production of speech. For example, in the case of the Mandarin tone 3 (low-dipping in terms of height and contour), the correlated head/neck motion during tone production should be signaled by a low-falling-rising movement. When present, these visual cues seem to be used by listeners during auditory-visual perception (Vatiktiotis-Bateson, Kroos, Kuratate, Munhall, & Pitermann, 2000).

It has been suggested that some lexical tones are easier to distinguish than others during audio-visual speech perception. For example, Mixdorff et al. (2005b) observed that Mandarin Chinese native speakers highly confused Mandarin tone 1 with tone 2 when they were asked to identify the tones in the devoiced-audio-visual condition. Additionally, they also found that tone 1 yielded the least correct responses, whereas tone 3 yielded the highest scores in the devoiced-audio-visual condition. Chen and Massaro (2008) also mentioned that the visual cues for tone 3 (neck movements) tended to be more pronounced than those for tone 2 and tone 4, and tone 4 tended to have the shortest duration of visual cues. Speakers presumably provided extra visual information when they used tone 3 through dipping their head or chin, which made tone 3 the easiest one to distinguish, while tone 2 and tone 4 were relatively hard to discriminate on the basis of visual information. Since facial motion might provide better clues for the identification of some tones than others, the effects of the visual modality may differ between tones.

Even though the main focus in this study is on the role of modality and speaking style in tone identification, the role of speaker and tone variation will be investigated as well, based on the considerations outlined above. However, these analyses should be considered as exploratory post hoc rather than as tests of predefined hypotheses.

2.1.4 The current study

As reviewed previously, the acquisition of tones for speakers of non-tonal languages is difficult. Tone languages are different from European languages, which tend to rely completely on phonological distinctions between vowels and consonants to distinguish word meanings (exceptions like Swedish, Serbo-Croatian, and a few Dutch Limburgian dialects notwithstanding). Tones serve to distinguish meanings at the lexical level (Yip, 2002) and tones can be viewed as phonemic distinctions that are attached to the syllable at a suprasegmental level. The main acoustic features of tones are fundamental frequency (Fo) (as the correlate of speech pitch), amplitude (intensity), and, to a lesser extent, duration (Kong, 1987; Francis et al., 2008; Hallé et al., 2004). In other words, tones are not only marked by melodic contrasts, but these tend to co-vary with other acoustic variables. The consensus is that Fo is the most dominant phonetic cue for Mandarin Chinese tones (Francis et al., 2008; Hallé et al., 2004; Kong, 1987). Based on Fo patterns (both height and contour) and the direction of pitch, tone 1 has been described as high level (5-5), tone 2 as mid-rising (or mid-high-rising; 3-5), tone 3 as low-dipping (also low-falling-rising or mid-falling-rising; 2-1-4), and tone 4 as high-falling (5-1) (Chao, 1930).

The goals of the current study are to explore to what extent people with no previous knowledge of tones can benefit in their ability to identify Mandarin

tones from (1) visual cues (i.e., whether or not a learner can see the speaker) and (2) the speaking style (i.e., whether or not the language input is transmitted in teaching style).

Given the beneficial effect visual cues are supposed to have for tone perception, the hypothesis we explore is: participants in the bimodal (audio-visual) condition will outperform participants in the unimodal (audio-only) condition, i.e. they will give more correct responses and have shorter reaction times. Similarly, we predict that participants exposed to the teaching style perform better than their counterparts who are exposed to the stimuli in natural style. Speaking style and modality most likely have independent effects, but there might also be interactions between the two. The combination of clear speech and visual cues might, for example, facilitate tonal learning more than each factor independently. In contrast (and in line with the absence of audiovisual superiority in clear speech conditions), visual information might be of little added value in a teaching style because of the clear audio signal. Based on the finding that audiovisual information is mostly beneficial in situations where the auditory signal is degraded, we conjecture that the difference between the audiovisual and audio-only condition is more pronounced in the natural speaking style condition, while in the teaching style condition, this difference is attenuated or perhaps even absent. The reasoning behind this, is that for tone-naïve listeners, normal speech presents less clear (e.g., "degraded") information about tone than teaching style.

We conducted a tone perception experiment, in which native speakers of Dutch were asked to distinguish Mandarin Chinese tones on the basis of audio or video materials. To account for variation between speakers and between tones, the experimental stimuli were recorded using four different speakers and four different tones. We report the proportion of correct responses and average reaction times. We use reaction times in addition to accuracy, because they have proven useful for indicating the degree of helpfulness of visual cues and teaching style (Chen, 2003; Schneider, Dogil, & Möbius, 2011) and to extend previous research that only reports the proportion of correct responses (Burnham et al., 2001; Chen & Massaro, 2008; Mixdorff et al., 2005a, 2005b).

2.2 Methodology

We employed a 2 (modality) × 2 (speaking style) design for our study. Participants were divided over two modality conditions (audio-visual vs. audio-only) and two speaking styles (natural vs. teaching). In addition to these between-subject factors, we also looked at the role of the speakers to which participants were exposed (two male and two female) and the variation in identification across tones (the four Mandarin tones; see previous section) as within-subject factors.

As dependent variables we recorded both the accuracy (whether a response was correct or not), and the reaction time (how long a participant took to respond) for each stimulus.

2.2.1 Participants

Eighty-six participants were recruited from the Tilburg University participant pool. The age of the participants ranged from 18 to 35 (M = 23, SD = 2.9). None of them had been previously exposed to tone languages. 72% of the participants were native speakers of Dutch, the remaining subjects were German, Italian, British, Spanish, Austrian, Indonesian, Bulgarian, and Turkish. They either received 0.5 study credits for their participation or a small token of appreciation. Participants were randomly assigned to one of the four conditions (video + teaching; video + natural; audio + teaching; audio + natural), while maintaining a balanced gender distribution in each group.

2.2.2 Stimuli

Stimulus construction

We constructed a word list with 10 Mandarin monosyllables (e.g., *ma*, *ying* …; selection based on Chen & Massaro, 2008 and Francis et al., 2008, see Appendix 1 for the complete list). Each of these syllables was chosen such that the four tones would generate four different meanings resulting in 40 (10 syllables × 4 tones) different existing words in Mandarin Chinese. Four adult native Mandarin-Chinese speakers (two male and two female) were asked to read out these words. All speakers were born and raised in China and had come to the Netherlands for their graduate studies. They have been in the Netherlands for less than three years. Speakers were instructed to produce the 40 words in two different scenarios in sequence: a natural mode ("pronounce these words as if you were talking to a Chinese speaker") and a teaching mode ("as if you were talking to someone who is not a Chinese speaker"). In both conditions, there were no other instructions or constraints imposed on the way they should produce the stimuli. There was a 20-minute break between the two recordings to avoid fatigue, with the recording of the natural stimuli preceding the recording of the teaching style stimuli.

We used Eye-catcher (version 3.5.1) and Windows Movie Maker (2012) to record the speakers' images and sounds. One of the advantages of Eye-catcher is that the camera is located behind the computer screen, which is convenient for unobtrusively capturing the full-frontal images of speakers' faces, similar to what listeners see in a face-to-face setting.

In total, 320 stimuli were produced; two sets of 160 video stimuli (10 syllables × 4 tones × 4 speakers) were generated, in teaching and in natural modes. These video clips were segmented into individual tokens, with each token containing one stimulus. In the final analysis, two problematic stimuli were discarded because one was edited too short, and another one was produced incorrectly by the speaker. We converted the video format from mp4 to avi using Freemake Video Converter (version 4.1.6) to ensure compatibility with E-Prime. Format Factory (version 3.9.5) was used to extract the sound from each video to generate the material for the audio-only conditions. This resulted in four types of experimental stimuli: video + teaching (VT); video + natural (VN); audio + teaching (AT); audio + natural (AN).

Pilot study

To ascertain the feasibility of the experimental task and the validity of the stimuli, 24 native Mandarin Chinese speakers were asked to identify the tones which were presented in the audio + natural condition (the supposedly most challenging condition). These speakers were born and raised in China. They were postgraduate students (aged 21-40) and had been staying in the Netherlands between four months and five years. They were asked to identify the tones, and their accuracy was 99.5%, indicating the validity of the stimuli and the feasibility of the task.

Stimulus characteristics

Acoustic and visual analyses were conducted to assess whether the differences between the two speaking styles (teaching and natural) were present in the experimental stimuli. We measured the mean duration and the average pitch of the two sets of experimental stimuli. In general, we expected that the duration of the teaching stimuli would be longer than the duration of the natural style stimuli. Similarly, we expected that tone patterns would be exaggerated in the hyperarticulated style. The tone fidelity, however, should not be impacted by the speaking style, since pitch is closely related to the lexical meaning of the word (Xu & Burnham, 2010).

We used Praat 6.0.33 (Boersma & Weenink, 2017) to measure the duration and average pitch of the experimental stimuli. A repeated-measures ANOVA, with tone and speaking style as the within-subject factors and speaker as between-subject factor, revealed that speaking style had a significant effect on the duration of the stimuli, F (1, 9) = 63.3, p < .001, η_p^2 = .876. In line with our expectation, the average duration of the stimuli in the teaching style (M = 0.54 ms, SE = 0.02) was longer than that in the natural style (M = 0.48 ms, SE = 0.02). The average pitch was *not* influenced by speaking style, with F (1, 9) = 2.09,

p = .183, η_p^2 = .188. Figure 2.1 provides an illustrative example of the difference between the two speaking styles for the four tones. The figure clearly shows the expected rising and falling patterns, which are more pronounced (especially in their duration) in the teaching style. Table 2.1 and Table 2.2 present the means and standard deviations (as well as confidence intervals) for the average pitch and duration. Speaker differences accounted for a decent amount of variation in the average duration of the experimental stimuli, F (3, 27) = 8.01, p = .001, η_p^2 = .471, and even more variation in average pitch, F (3, 27) = 28.7, p < .001, η_p^2 = .761 (Table 2.1). Tones accounted for a large amount of variation between the two speaking styles in duration: F (3, 27) = 399.8, p < .001, η_p^2 = .978 and in pitch: F (3, 27) = 66.4, p < .001, η_p^2 = .881(Table 2.2). There was a significant interaction between tone and speaking style in terms of duration: F (3, 27) = 9.01, p < .001, η_p^2 = .50. However, no significant interactions were found between tone and speaking style in terms of average pitch: F (3, 27) = 0.59, p > .05, η_p^2 = .63. Thus, our global acoustic analysis reveals strong effects of individual speakers and tones, while small differences between the two speaking styles also emerge.

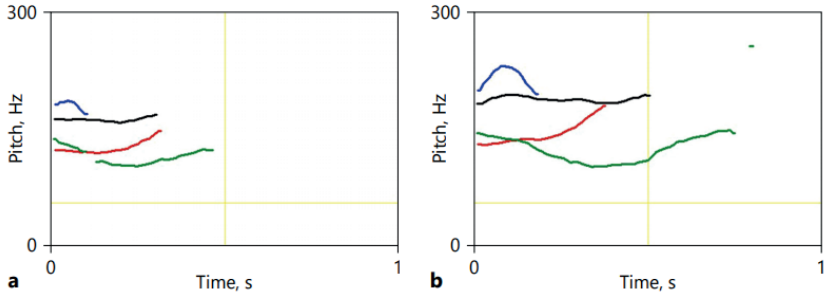


Figure 2.1. Plots of tone contours for natural (**a**) and teaching style (**b**). Figure based on one male speaker producing syllable /ma/.To illustrate differences in duration and pitch between the two speaking styles, the scale of the *x* and *y* axis is kept identical: time (0-1s) on the *x* axis and pitch (0-300 Hz) on the *y* axis. Tone 1, black; tone 2, red; tone 3, green; tone 4, blue.

Speaker	Tone	Mean		SE		95% CI			
						lower bound		upper bound	
		N	T	N	T	N	T	N	T
1	1	.434	.564	.036	.032	.354	.493	.514	.635
	2	.509	.633	.039	.029	.421	.568	.597	.698
	3	.663	.815	.032	.033	.590	.739	.736	.891
	4	.373	.375	.028	.021	.310	.328	.436	.422
2	1	.454	.567	.029	.036	.389	.485	.519	.649
	2	.458	.510	.026	.035	.399	.430	.517	.590
	3	.701	.848	.036	.037	.621	.765	.781	.931
	4	.395	.419	.026	.024	.337	.364	.453	.474
3	1	.429	.436	.026	.031	.370	.366	.488	.506
	2	.507	.519	.024	.027	.452	.459	.562	.579
	3	.588	.640	.033	.029	.514	.575	.662	.705
	4	.349	.360	.027	.028	.288	.298	.410	.422
4	1	.413	.431	.025	.027	.357	.370	.469	.492
	2	.490	.536	.031	.027	.419	.474	.561	.598
	3	.590	.687	.019	.029	.546	.622	.634	.752
	4	.336	.343	.024	.031	.282	.273	.390	.413

Table 2.1. Descriptive statistics for duration (seprately by speaker, tone and style) (s). Note that SE represents standard error; CI represents confidence interval; N represents natural style; T represents teaching style.

Speaker	Tone	Mean		SE		95% CI			
						lower bound		upper bound	
		N	T	N	T	N	T	N	T
1	1	232.662	252.691	5.253	4.534	220.778	242.434	244.546	262.948
	2	199.245	208.177	5.039	7.435	187.847	191.357	210.643	224.997
	3	168.108	187.620	9.136	12.746	147.441	158.787	188.775	216.453
	4	295.655	322.456	12.627	10.317	267.091	299.118	324.219	345.794
2	1	172.967	195.988	5.805	5.436	159.834	183.690	186.100	208.286
	2	149.407	160.717	6.898	3.911	133.802	151.869	165.012	169.565
	3	133.542	138.875	5.683	4.836	120.686	127.935	146.398	149.815
	4	232.851	221.915	24.496	13.603	177.436	191.142	288.266	252.688
3	1	156.752	138.275	18.896	3.102	114.006	131.258	199.498	145.292
	2	114.581	135.800	8.987	11.564	94.252	109.641	134.910	161.959
	3	170.019	168.692	36.792	26.611	86.791	108.495	253.247	228.889
	4	253.851	222.173	36.442	30.937	171.413	152.188	336.289	292.158
4	1	305.976	323.645	3.548	5.989	297.950	310.098	314.002	337.192
	2	229.345	241.122	12.771	9.022	200.455	220.712	258.235	261.532
	3	200.173	194.949	10.446	5.479	176.543	182.554	223.803	207.344
	4	346.520	348.839	20.047	13.421	301.171	318.480	391.869	379.198

Table 2.2. Descriptive statistics for average pitch (separated by speaker, tone, and style) (Hz). Note that SE represents standard error; CI represents confidence interval; N represents natural style; T represents teaching style.

For the visual analyses, we expected that the hyperarticulated action (teaching style) results in more facial movements as compared to natural style. We used Flow Analyzer[4] to track the amount of movement present in the video as an estimate of the magnitude of movements. In this case, the total amount of motion was measured for each speaker, in both speaking styles, for each of the syllable/tone combinations. A repeated-measures ANOVA showed that speaking style had a main effect on the total amount of motion, F (1, 9) = 115, $p < .001$, $\eta_p^2 = .928$. In teaching style (M = 0.25, SE = 0.01), speakers tended to use more visual cues than in the natural style (M = 0.15, SE = 0.003), which

[4]Flow Analyzer is a piece of software, based on Optical Flow Analysis, for extracting motion from 2D video sequences. Optical flow computes pixel displacements between consecutive frames in the video.

is in line with the idea of hyperarticulation. Individual speakers also differed significantly in their amount of movement, F (3, 27) = 19.56, p < .001, η_p^2 = .685. Pairwise comparison (using Bonferroni adjustment) showed that Speaker 1 and speaker 3 provided the most visual movement information. There is no difference between speaker 1 (M = 0.27, SE = 0.023) and speaker 3 (M = 0.21, SE = 0.008). Speaker 4 (M = 0.18, SE = 0.003) moved significantly less than speaker 1 and speaker 3, p< .02. Speaker 2 (M = 0.13, SE = 0.009) signaled the least visual information. The different tones did not affect the amount of movement, F (3, 27) = 2.27, p = .103, η_p^2 = .202. On average, there was no significant difference in motion between natural and teaching speaking style for each separate tone: F (3, 27) = 2.79, p > .05, η_p^2 = .237.

Flow Analyzer can also measure the motions displayed in the horizontal direction (x) and the vertical direction (y), which can give a clearer picture of the directionality or type of motions among different tones. Table 2.3 provides the amount of movement in the x and y direction and shows for example, that tone 1 has the least amount of vertical movement and the most horizontal movement, which is in line with a level tone. Similarly, for tone 2, tone 3 and tone 4, there is more vertical than horizontal motion.

Tone	Mean		SE		95% CI			
					lower bound	upper bound	lower bound	upper bound
	x	*y*	*x*	*y*	*x*	*x*	*y*	*y*
1	.037	.029	.005	.001	.026	.047	.026	.032
2	.028	.039	.003	.002	.023	.034	.035	.043
3	.025	.047	.001	.003	.022	.029	.041	.053
4	.033	.047	.001	.003	.030	.036	.040	.055

Table 2.3. Descriptive statistics for the amount of facial movements on x and y axes for the different tones (pixels per frame). Note that SE represents standard error and CI represents confidence interval.

In sum, the results of the acoustic and visual analyses show important differences and interactions in the stimuli between speaking styles in both the auditory and visual domains, illustrating the need to test the influence of speaking style and visual information (modality) on Mandarin tone identification by naïve listeners.

2.2.3 Procedure

All sessions were conducted in a sound-attenuated room. E-prime (version 2.0; Zuccolotto, Roush, Eschman, & Schneider, 2012) was used to set up and run the experiment. The full procedure consisted of three blocks: instruction, practice trials, and test trials. Before the experiment started, participants were asked to fill out a questionnaire that assessed their language background. After that, a brief instruction about Mandarin Chinese tones was first displayed on the screen (see Figure 2.2 for a screenshot): "There are four tones in Mandarin Chinese: the first tone is a High-Level tone, symbolized as " ˉ ", the second tone is a Mid-Rising tone, symbolized as "ˊ", the third tone is a Low-Dipping tone, symbolized as " ˇ ", and the fourth tone is a High-Falling tone, symbolized as " ˋ ".

The task of the participants was to identify the tones they perceived from the speakers. Three practice trials were included to allow participants to get familiar with the testing procedure and the stimuli. After the practice trials, the experiment leader checked with the participants to make sure they fully understood the concept of tones (in particular the symbols) and the task. Finally, 160 testing stimuli (video/audio) were presented in randomized order for each participant (operated by E-Prime). The time for participants to give responses was 10 seconds. Participants received feedback: "good job" or "incorrect" depending on the correctness of their response, or, "no response", if the participants had not reacted within the given 10 seconds. In order to motivate the participants to do their best, a special programming code was implemented in the experimental procedure: if the participants gave ten correct responses consecutively, the experiment would stop[5].

Participants wore headsets, and were seated directly in front of the PC running the experiment. All stimuli were presented at a comfortable hearing level. The participants were instructed to press the designated keys with the corresponding tone symbols (" ˉ ", "ˊ", " ˇ ", " ˋ ", see Figure 2.3) on them as accurately and as quickly as possible after they made their decisions. Their responses and reaction times were recorded automatically by E-prime.

[5]In total, 22 participants finished their experiments in advance (they gave ten correct responses consecutively). The distributed numbers in different conditions are: 8 in video + teaching; 7 in audio + teaching; 6 in video + natural; 1 in audio + natural

Welcome to the experiment.

You will see several video clips in which you can hear Chinese people speak Chinese words. After the video, four tone symbols will be displayed.

Four tones in Chinese:

Tone 1: high-level " ˉ "

Tone 2: mid-rising " ˊ "

Tone 3: low-dipping " ˇ "

Tone 4: high-falling " ˋ "

PRESS THE SPACEBAR TO CONTINUE...

Figure 2.2. Screenshot of a brief introduction of Mandarin Chinese tones (in video conditions)

Figure 2.3. Picture of the designated keys with tone symbols on them

2.3 Results

The goal of this study was to examine to what extent modality (audio-visual vs. audio-only information) and speaking style (natural vs. teaching mode) affect the perception of Mandarin Chinese tones in naïve listeners. For each stimulus, we recorded accuracy (whether a response was correct or not) and reaction time (how long a participant took to respond). The proportion of correct responses (accuracy) and latency (the time elapsed between the onset of the stimulus and the onset of the response, expressed in milliseconds) will be presented as dependent variables. For each dependent variable, a repeated-measures ANOVA was carried out with modality and speaking style as between-subject factors, and speaker and tone as within-subject factors.

This section contains two parts: first, we present a general picture with respect to the effects of the between-subject variables, followed by the results regarding the within-subject variables. The interactions between variables are also described at the end of each part.

2.3.1 The effects of modality and speaking style

Figure 2.4 and Figure 2.5 present the effect of modality and speaking style on accuracy and reaction time. The first analysis examined whether communication modality and speaking style affected the accuracy of Mandarin tone perception.

As Figure 2.4 shows, overall, the video condition (M = 49.1%, SE = 0.02) resulted in higher accuracy scores than the audio condition (M = 42.1%, SE = 0.02), and the difference between them was statistically significant (F (1, 76) = 6.74, p = .011, η_p^2 = .08). This is in line with the hypothesis that the availability of visual cues along with auditory information should benefit people who have no previous knowledge of Mandarin Chinese tones. No significant effects, however, were observed in terms of reaction times between the two modalities: M = 1048 ms, SE = 71.5 for video; M = 1197 ms, SE = 77.1 for audio; F (1, 76) = 2.00, p = .161, η_p^2 = .26.

As shown in Figure 2.5, participants exposed to teaching style were better at tone identification (M = 49.6%, SE = 0.02) and responded faster to the stimuli (M = 1062 ms, SE = 75.3) than participants exposed to natural speaking style materials (M = 44.7%, SE = 0.02; M = 1183 ms, SE = 73.4). However, these differences were not significant (F (1, 76) = 0.49, p = .484, η_p^2 = .01 for accuracy, and F (1, 76) = 1.31, p = .256, η_p^2 = .02 for reaction time). These results therefore reject our hypothesis that speaking style would have a main effect on the speed or accuracy of Mandarin tone perception. Furthermore, as shown in Figure 2.6, no significant interaction was observed between modality and speaking style, neither for accuracy nor for reaction time (F (1.76) = 1.39, p = .243; F (1.76) = 0.23, p = .631).

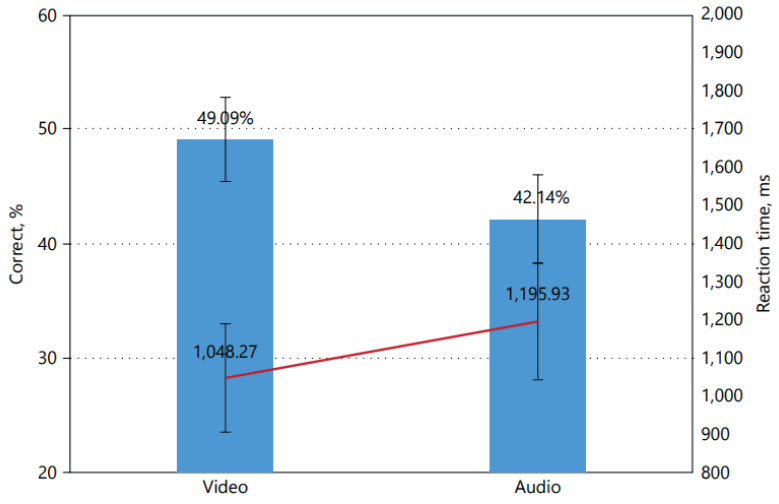


Figure 2.4. Illustration of the participants' performance in video (audio-visual) and audio (audio-only) conditions: bars represent the percentage of correct responses (accuracy), and lines represent the average reaction time

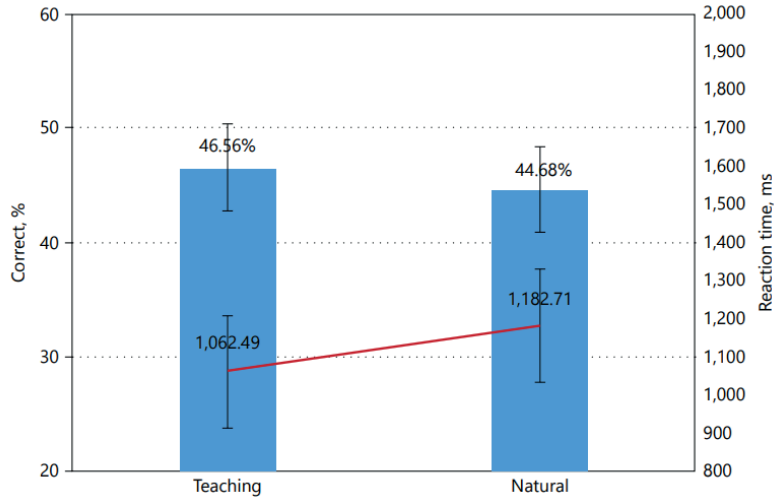


Figure 2.5. Illustration of the participants' performance in teaching and natural styles: bars represent the percentage of correct responses (accuracy), and lines represent the average reaction time

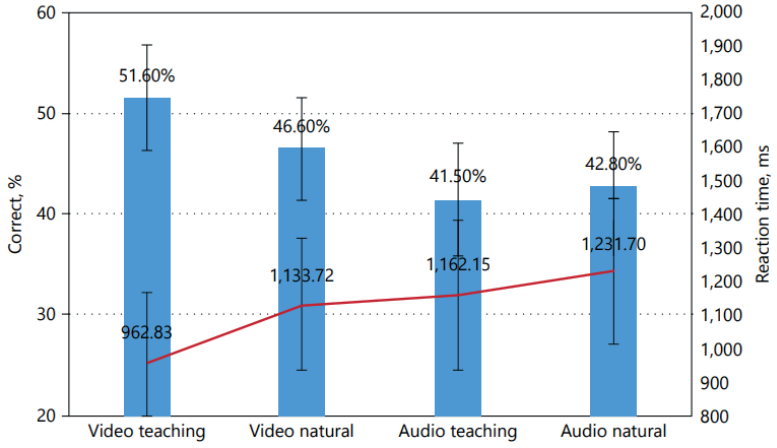


Figure 2.6. Illustration of the interaction between modality (video/audio) and speaking style (teaching/natural): bars represent the percentage of correct responses (accuracy), and lines represent the average reaction time

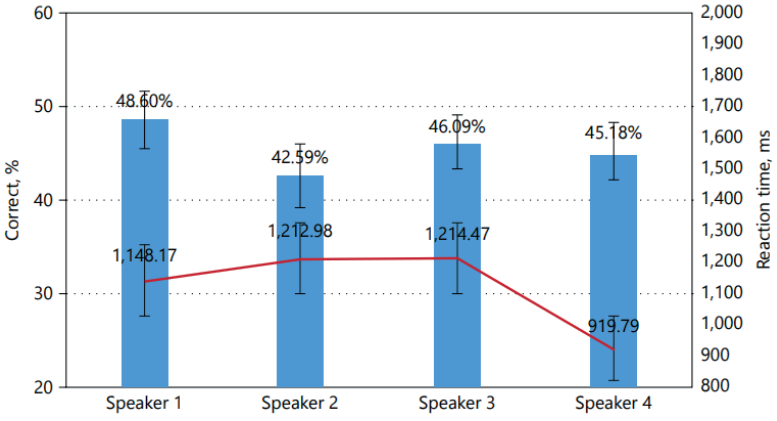


Figure 2.7. Illustration of participants' reaction to different speakers: bars represent the percentage of correct responses (accuracy), and lines represent the average reaction time

2.3.2 The effects of speaker and tone

In the second part, we investigated whether individual speaker variation and differences in tones affected the speed and accuracy of Mandarin tone perception.

Figure 2.7 illustrates the differences between speakers. There is a significant main effect on accuracy, F (3, 228) = 7.05, p < .001, η_p^2 = .09. Participants gave the most correct responses when the tones were produced by speaker 1 (M = 48.6%, SE = 0.02), and the least correct responses were given for speaker 2 (M = 42.6%, SE = 0.02). Pairwise comparisons (Bonferroni corrected for multiple comparisons) showed that only the difference between speaker 1 and speaker 2 was significant (D = 0.06, p < .001), while the differences between those speakers and the other two were not. There was also a significant main effect for speaker on average reaction time, F (3, 228) = 43.9, p < .001, η_p^2 = .366, which was mainly due to the participants' swift responses (M = 920 ms, SE = 52.4) to the stimuli produced by speaker 4.

The scenario becomes more complicated when we consider interactions between the independent variables. There was a significant interaction between speaker and modality on both accuracy and reaction time (see Figure 2.8): F (3, 228) = 10.62, p < .001, η_p^2 = .12 in terms of accuracy, and F (3, 228) = 5.33, p = .001, η_p^2 = .07 in terms of reaction time. Another significant interaction was observed between speaker and speaking style (see Figure 2.9), the corresponding values are F (3, 228) = 4.94, p = .002, η_p^2 = .06 for accuracy and F (3, 228) = 9.36, p < .001, η_p^2 = .11 for reaction time. Finally, there were significant three-way interactions of speaker, modality and speaking style on both accuracy and reaction time: F (3, 228) = 5.04, p = .002, η_p^2 = .06 and F (3, 228) = 4.72, p = .003, η_p^2 = .06, respectively. These interactions suggest that individual speaker characteristics should be considered in these kind of studies, because of their intricate interplay with factors such as speaking style and modality.

Figure 2.8 depicts the interaction between speaker and modality (audio-visual vs. audio-only). For both accuracy and reaction time, the conclusion that the audio-visual condition yields better results than the audio-only condition actually depends on which speaker produced the material. As shown in the graph, it clearly benefited participants more when they saw *and* heard speaker 1 (female) and speaker 3 (male) as compared to only hearing them. There were no such obvious differences with speakers 2 (male) and 4 (female), though they also showed better performance in the audio-visual condition.

Figure 2.9 presents the interaction between speaker and speaking style (teaching vs. natural). As shown in the graph, for some speakers (speaker 1 and speaker 3), the teaching style helped the participants give more correct responses and react faster compared to the natural style. However, participants did not benefit from the teaching style with speakers 2 and 4. When speaker 4

produced tone in the teaching style, it actually took participants more time to react.

Figure 2.10 illustrates the results for the four different tones. Given that there are four tones, the statistical chance level for a correct response is at 25%. The results show that the overall performance levels were above chance for each of the tones. Regardless of condition, tone had a striking main effect on the percentage of correct responses and on average reaction time, $F(3, 228) = 52.1$, $p < .001$, $\eta_p^2 = .41$; and $F(3, 228) = 13.98$, $p < .001$, $\eta_p^2 = .16$, respectively. A follow-up analysis with pairwise comparison (with Bonferroni correction for multiple comparisons) shows that participants generally performed best on tone 3: they gave more correct responses ($M = 58.6\%$, $SE = 0.02$) and were faster ($M = 1026$ ms, $SE = 48.3$) than when they heard the other tones. Tone 4 was the most challenging one for participants to identify ($M = 29.5\%$, $SE = 0.02$; $M = 1220$ ms, $SE = 60.2$).

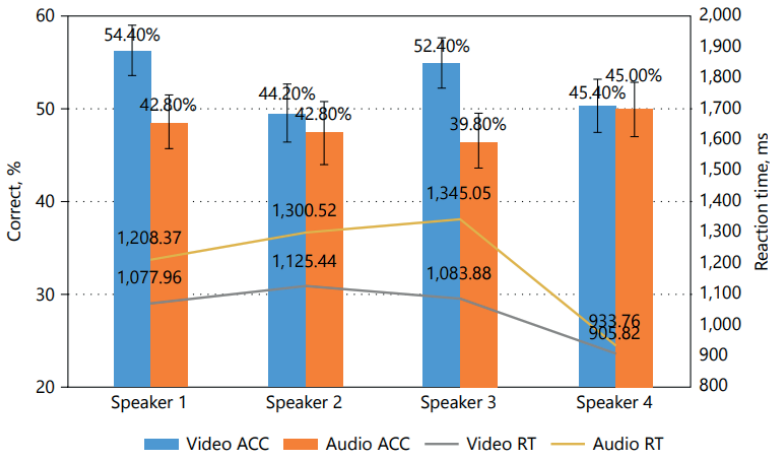


Figure 2.8. Illustration of the interaction between speaker and modality (video/audio): bars represent the percentage of correct responses (accuracy), and lines represent the average reaction time

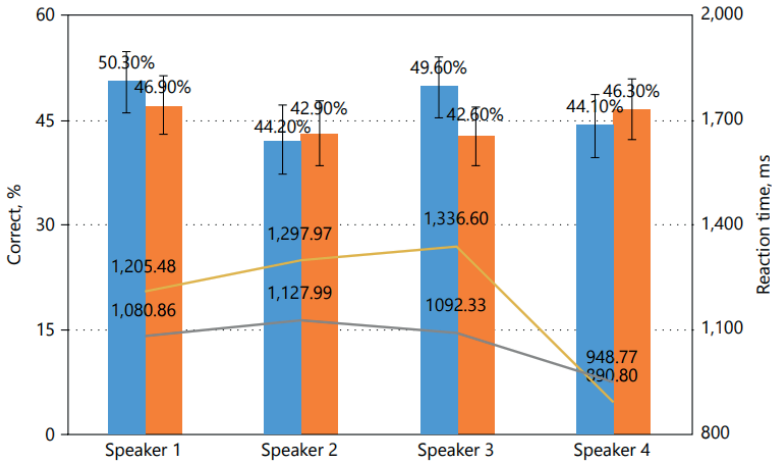


Figure 2.9. Illustration of the interaction between speaker and speaking style (teaching/natural): bars represent accuracy, and lines represent the average reaction time

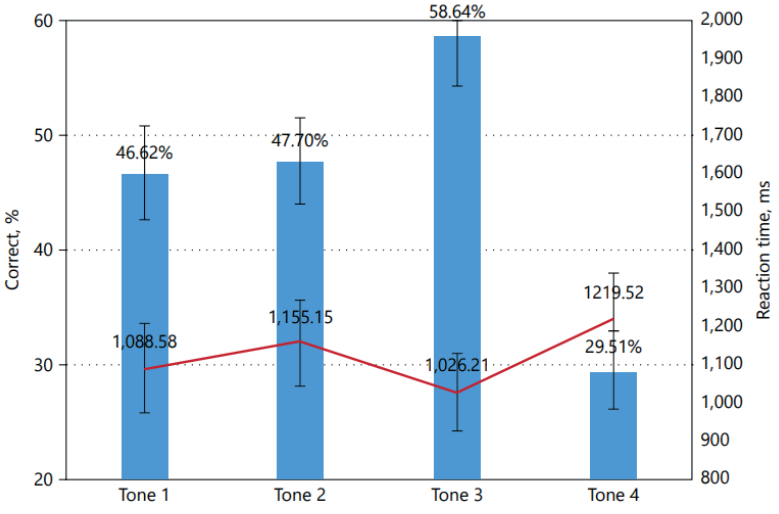


Figure 2.10. Illustration of participants' reaction to different tones: bars represent the accuracy, and lines represent the average reaction time

The corresponding tonal confusion matrix is displayed in Table 2.4: it shows that tone 3 was indeed the least confusing tone (68.9%) for non-native listeners. The falling-tone tone 4 was most commonly misidentified as a high-level tone 1 (26.5 %), though the confusions are not necessarily symmetrical: tone 1 was mostly confused with mid-rising tone 2 (20.3 %), rather than with tone 4.

	Presented Tones				Total
	1	2	3	4	
Missing	0.17	0.10	0.12	0.10	0.13
Responded tones					
1	**59.58**	13.03	5.10	26.52	26.16
2	20.27	**59.87**	13.12	18.14	27.75
3	7.51	17.87	**68.86**	7.89	25.67
4	12.46	9.13	12.80	**47.35**	20.30
Total count	4,020	3,922	4,001	3,892	15,835

Table 2.4. Confusion matrix for tone (percentage correct). Rows indicate responded tones from participants, columns present tones from stimuli, and figures correspond to percent. Bold figures indicate correct tone identification.

There was a significant interaction between speaker and tone for correct responses (F (9, 684) = 4.25, $p < .001$, $\eta_p^2 = .05$), but not for reaction times (F (9, 684) = 1.63, $p = .102$, $\eta_p^2 = .02$).

With regard to reaction time, there was indeed a significant interaction between tone and speaking style (see Figure 2.11): F (3, 228) = 7.30, $p < .001$, $\eta_p^2 = .09$. For each tone in the two styles, participants responded faster in the teaching condition than in the natural style (M = 1062 ms, SE = 75.3 vs. M = 1183 ms, SE = 73.4). This tells us that the speed of people's reactions to the tone depends on the speaking style. Teaching style may not help listeners give a more correct response, but it does speed up their decisions. From Figure 2.11, we can also see that teaching style helps listeners to identify some certain tones faster: listeners were only significantly faster when they heard tone 2 or tone 4, while the difference between tone 1 and tone 3 was not significant.

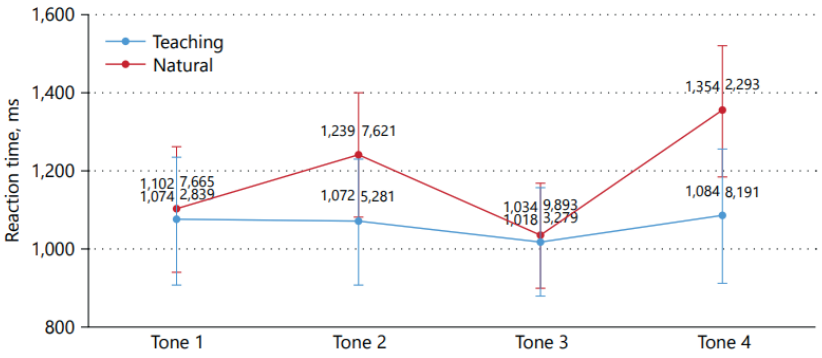


Figure 2.11. Illustration of the interaction between tone and speaking style (teaching/natural) in terms of reaction time. Teaching style, blue; natural style, red.

2.4 Discussion and conclusion

The main goals of the current study were to explore to what extent tone-naïve listeners of Mandarin Chinese are able to learn to identify tones in isolated words, and whether they can benefit (1) from visual cues that speakers display in their facial movements and (2) from a speaking style that resembles teacher talk.

Our finding that tone-naïve listeners were better able to identify tones when they saw the speakers than when they only heard them supports the hypothesis that visual information plays a facilitating role in learning to identify Mandarin tones. This suggests that there is perceptually-salient visual information that aids in the classification of Mandarin tones. These visual cues, including movements of the head, neck, mouth and lips, are observable and used by non-native listeners as features to discriminate between the different tones. Interestingly, many previous studies were not able to find significant differences between audio-only and audio-visual conditions in native speakers except under difficult listening conditions (e.g., Sumby & Pollack, 1954; Smith & Burnham, 2012) or degraded stimuli (Burnham et al., 2001; Mixdorff & Charnvivit, 2004). One explanation for this discrepancy between previous and current results could be that earlier studies on visual information used native speakers as participants. In our pretest, we also observed that native participants in the audio-only condition already performed at ceiling (see section 2.2.2). In earlier studies, these may have obscured possible beneficial effects of the visual modality. In contrast to these previous findings, we did find a difference between audio-only and audio-visual conditions with "clear" stimuli (i.e., no noise or degraded signal) for non-native listeners in Mandarin Chinese tone

perception. In that respect, our study is in line with studies that have shown that visual information is available to, and used by, language learners during speech perception (Burnham et al., 2001, 2002; Reid et al., 2015).

In contrast, and contrary to our expectations, we did not find a significant main effect of speaking style on identification accuracy or reaction time. Stimuli produced in a teaching style were not identified better than those produced in a natural style, neither in terms of accuracy nor in terms of reaction time. It should be noted, though, that there was a significant two-way interaction between tone and speaking style on average reaction time. This interaction reveals that the speed with which listeners can identify specific tones is influenced by speaking style for selected tones. More specifically, listeners were faster for tone 2 and tone 4 in teaching style. This is probably due to the fact that the tone contours for the rising tone 2 and falling tone 4 are hyperarticulated or exaggerated to a greater degree (see also Kim & Davis, 2001) in the teaching style, which eases the task for listeners (Rosenblum & Fowler, 1991). We found no differences for tones 1 and 3, possibly due to the intrinsic acoustic features of these two tones. Tone 1 is a level tone, for which it is hard to see how hyperarticulation could be beneficial. Tone 3 is the easiest one for listeners to distinguish due to its unique tone shape (first falling, then rising) and duration. In addition, given the set-up of our stimulus recording (with elicited productions of isolated words), it could have been the case that our natural condition already represents rather "clear" speech, and in that sense it is not representative of the reduced speech samples one often observes in more spontaneous data, so that this tone may already have been comparatively easy for listeners to distinguish.

We followed up our investigation of speaking style and modality with a (post hoc) analysis of speaker effects. Specifically, we assessed whether some speakers produce tones that are easier to identify than the tones of other speakers. We found that participants indeed gave more correct responses to the stimuli produced by speaker 1, and reacted faster to those produced by speaker 4. Speaker variation in general appears to have a significant influence on Mandarin tone identification: speaker 1 is an easier model for the participants to recognize which results from the combined effect of acoustic and visual information she provided. For example, speaker 1 provided more visual motion than speaker 2 and speaker 4 did (even when the difference between speaker 1 and speaker 3 was not significant in this respect). This difference is visible not only in the numbers, but also in a more qualitative analysis of her recordings: obvious visual motions are present when she produced tone 1 by horizontally moving her head and neck, while the other speakers barely used such motions when they produced the same tone. In the audio-only conditions, the words spoken by speaker 1 had a longer duration on average than those of the other speakers, which could also have been beneficial for tone identification. In addition, given that speakers 1 and 4 are both female, it would also be interesting

to explore in greater detail whether female speakers are easier models to learn from, and, if so, what kinds of characteristics are responsible for this beneficial effect. Our study does not allow us to attach much importance to the gender differences, however, due to the limited number of speakers.

Finally, apart from the effects of modality and speaking style, inherent features of the tones appear to be main contributors to differences in tone perception. In other words, it is much more important *which* tone the listeners hear than *how* they hear it. We found that tone 3 was the easiest one for the listeners to identify, while tone 4 was the most difficult one. This is possibly due to their specific temporal characteristics – tone 3 has the longest duration and two intensity peaks, while tone 4 has the shortest duration, and only one intensity peak. This finding is not entirely in line with previous studies: while tone 3 has indeed been found the easiest to identify, due to the longest vowel duration, tone 4 has not been clearly found to be the most difficult one to recognize (Mixdorff et al., 2005b; Blicher, Diehl, & Cohen, 1990; Fu & Zeng, 2000).

So, our findings imply that teaching Mandarin tones might benefit from pointing out to learners the information that visual cues can contribute. In a teaching context, teachers should consider using their facial expressions while talking to students/learners, while learners/students should be trained to consciously attend to visual information. As for online learners, using video is likely to be more effective than using audio-only material.

Further research could reveal the underlying causes of the superiority of audio-visual presentation over visual-only stimuli. For instance, in order to figure out which information is more influential (the auditory information or the visual cues), researchers could measure the effect of discrepant visual information with auditory information during tone perception, similar to what has been done to test the McGurk effect (see Chapter 4). When listeners perceive conflicting auditory and visual information, the question is how they will weigh these two cues. For instance, a participant could be presented with an audiovisual presentation of a Mandarin low-dipping tone 3, either accompanied with a video of a speaker who produces lowering head and neck movement (which would be the most natural visual expression to accompany such a tone), or with an incongruent video of a speaker who produces a raising head and neck (that usually accompanies tone 2). In the latter case, the question then arises as to whether a participant would ignore one of the two sources of information, or integrate them in a way that may lead to a perception that is different from a unimodal presentation.

Overall, learning Mandarin tones may be facilitated by being aware of the aid provided by audio-visual information as well as by the potential benefit in clear speech as exemplified in teacher talk, although the contribution of both factors depends on the specific tones and speakers in question. Irrespective

of different contributions of each modality, some speakers/teachers will be easier models for learners to learn from than others. In addition, it is clear that the individual tones differ in learnability; this too should be taken on board in curriculum design. We hope our findings aid second language learners of Mandarin and will inspire further research on Mandarin tone learning.

Chapter 3

Chapter 3

Mandarin Tone Identification by Tone-naïve Musicians and Non-musicians in Auditory-visual and Auditory-only Conditions

Abstract. A considerable number of studies have shown that musical ability has a positive effect on language processing. Extending this body of work, this study investigates the effects of musicality and modality on Mandarin tone identification in tone-naïve participants. To examine the effects of visual information in speech, Mandarin tones were presented in auditory-only or auditory-visual modalities to participants with or without musical experience. The Goldsmith Musicality Index was used to assess the musical aptitude of the participants. Overall, musicians outperformed non-musicians in the tone identification task in both auditory-visual and auditory-only conditions. Both groups identified tones more accurately in the auditory-visual condition than in the auditory-only condition. In addition, performance differed by tone: musicality holds its main effect on each level of tone; while the influence of modality differs for individual tones; the identification of tone 3 (a low-falling-rising) proved to be the easiest, while tone 4 (a high-falling tone) was the most difficult to identify for all participants. Out of all the musical skills measured by the Goldsmith Musicality Index, the amount of musical training was the only predictor that had an impact on the accuracy of Mandarin tone perception. These findings suggest that learning to perceive Mandarin tones benefits from musical expertise, and visual information can facilitate Mandarin tone identification, but mainly for tone-naïve non-musicians.*

*This chapter is based on: Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2019). Mandarin tone identification by tone-naïve musicians and non-musicians in auditory-visual and auditory-only conditions. *Frontiers in Communication*, 4, 70. https://doi.org/10.3389/fcomm.2019.00070

3.1 Introduction

More than half of the languages (60% -70%) spoken in the world are so-called tone languages (Yip, 2002). Of these, Mandarin Chinese is spoken by the largest population by far (total users in all countries in 2015: 1,107,162,230[6]). Learning to identify Mandarin tones is difficult for speakers of non-tonal languages. Unlike most European languages, which rely primarily on phonological distinctions between consonants and vowels to distinguish word meanings, tone languages, such as Mandarin Chinese, additionally use tones to distinguish meanings at the lexical level. Marked by fundamental frequency (F0), pitch patterns and intrasegmental prosody, Mandarin Chinese has four main distinctive tones, conventionally numbered 1– 4: tone 1: high-level (5-5[7]); tone 2: mid-rising (or mid-high-rising; 3-5); tone 3: low-dipping (also low-falling-rising or mid-falling-rising; 2-1-4); and tone 4: high-falling (5-1) (Chao, 1930). Although tonal movement tends to correlate with other acoustic variables, the consensus is that F0 (as the correlate of perceived pitch) is the dominant acoustic feature for Mandarin Chinese tones (Tseng, 1981). Given the ubiquity of tonal languages and their increasing economic importance (Maddieson et al., 2013), identifying factors that promote efficient learning of Mandarin tones has attracted considerable scholarly attention (for example, Hao, 2012; So & Best, 2010). In the current study, we focus on two factors which may contribute to Mandarin tone perception: musical ability (comparing musicians and non-musicians), and modality (comparing auditory-visual stimuli with auditory-only stimuli).

3.1.1 Tone perception and musical ability

Musical ability has been shown to be an important factor in many aspects of language learning. Neuropsychological as well as behavioral studies have revealed that musical expertise positively influences aspects of speech processing such as lexical pitch (Alexander, Wong, & Bradlow, 2005; Delogu, Lampis, & Belardinelli, 2006, 2010; Ong, Burnham, Escudero, & Stevens, 2017), sentence intonation, and perceiving the metric structure of words (Marie et al., 2011). Both the perception of native (Schön et al., 2004) and foreign language speech (Marques et al., 2007) have been reported to benefit from musical experience (Marie et al., 2011; Milovanov et al., 2008; Milovanov et al., 2010). The current study aims to explore whether musical expertise also helps tone-naïve listeners to correctly identify Mandarin Chinese tones.

[6]Ibid., p. 4.
[7]Ibid., p. 4.

It is not surprising that musical expertise facilitates speech perception, since music and speech are similar in several ways (Besson, Chobert, & Marie, 2011; Patel, 2010). For one thing, music and speech are both complex auditory signals based on similar acoustic parameters: both pitch and duration contribute to the melodic and rhythmic aspects of music and to the linguistic functions of speech (Chobert & Besson, 2013). In addition, music and speech processing both require attention, memory and similar sensorimotor abilities. Furthermore, recent insights suggest that processing music and language use closely related neurocognitive systems. Although the dominant view has been that language and music processing were located in different hemispheres of the brain (left for language and right for music), an increasing number of studies have found that there is a functional overlap in the brain networks that process acoustical features used in both speech and music (Besson, Schön, Moreno, Santos, & Magne, 2007; Mok & Zuo, 2012; Patel, 2010; Wong, Skoe, Russo, Dees, & Kraus, 2007). For example, Tillmann, Burnham, Nguyen, Grimault, Gosselin and Peretz (2011) found that deficits in musical processing in non-tone language speakers with amusia were associated with deficits in lexical tone processing. Besides, musical training appears to drive adaptive plasticity in speech processing networks (Milovanov & Tervaniemi, 2011) and there is a music training transfer between music and acoustic processing in speech, such as frequency and duration (Besson et al., 2011). In line with the findings above, one would expect musicians to exhibit superior performance on pitch processing, and, as a result, being better at learning to discriminate tones.

The unfamiliarity with tone in many Western speakers makes tone languages ideally suited to examine the influence of musical experience on language acquisition (Marie et al., 2011). Previous studies have shown that musicians are more sensitive to subtle pitch variations in speech than non-musicians (e.g., Micheyl, Delhommeau, Perrot, & Oxenham, 2006; Schön et al., 2004). Furthermore, Burnham, Brooker and Reid (2015) showed that musicians and absolute pitch musicians were not only better overall in discriminating tones in speech and non-speech, but also less susceptible to the linguistic context effect. Behavioral studies clearly provided evidence that lexical tone perception benefits from musical expertise. For example, a relevant study by Gottfried and Riester (2000) showed that tone-naïve English music majors identified the four Mandarin tones better than non-musicians, and that musicians were also better at producing the Mandarin tones as compared to non-musicians. Furthermore, music majors performed better than non-musicians in pitch glide identification, and were more accurate in their identification of both intact and silent-center Mandarin syllable tones (Gottfried, Staby, & Ziemer, 2004; see also Alexander et al., 2005 for similar results). In another study, by using intact and acoustically modified syllables (silence-center syllables and onset-only syllables) of the four Mandarin tones produced by multiple speakers, Lee and Hung (2008) assessed

the difference in performance in Mandarin tone identification between English musicians (with 15 years of musical training on average, without absolute pitch abilities) and non-musicians. They found that musicians processed pitch contours better than non-musicians and concluded that (extensive) musical training facilitated lexical tone identification, although the extent to which musical ability facilitated tone perception varied as a function of the tone in question and the type of acoustic input. Specifically, the advantage of the musicians (in accuracy and reaction time) in identifying Mandarin tones decreased when the acoustic information was reduced (from intact syllables to silent-center and onset-only syllables); and musical background mainly benefited the identification of tones 1, 2, and 4. Taken together, these studies show that musicians consistently outperform non-musicians in the area of lexical tone processing of non-tone language speakers.

Much previous related research, such as the studies mentioned above, has focused on comparing musicians and non-musicians with regards to cognition, behavior and brain structure/function (Aheadi, Dixon, & Glover, 2010; Gaser & Schlaug, 2003; Hassler & Gupta, 1993; Koelsch, Gunter, Friederici, & Schröger, 2000). These earlier studies have usually compared two groups of participants, musicians and non-musicians, based on musical abilities conferred by musical training/education or based on the skill/level of playing musical instruments. While these criteria suffice to distinguish two different groups in general, they fail to provide insights into which aspects of musical ability contribute to the improved tone perception in musicians, since an individual's musicianship status is not a unitary construct, but comprises multiple abilities, such as singing ability, perceptual abilities, and duration of training. In the current study, we use the Goldsmiths Musical Sophistication Index (Gold-MSI) (Müllensiefen et al., 2014) as a tool to provide a more fine-grained analysis of the reported musical abilities of participants. With this, we aim to better understand the differences between musicians and non-musicians and relate these differences in specific musical abilities to different performance in tone perception.

3.1.2 Tone perception and visual information

The extent to which musicians outperform non-musicians in tone perception could be mediated by other factors, such as the presence of visual information, which has been shown to facilitate speech perception (e.g., Hirata & Kelly, 2010; Sueyoshi & Hardison, 2005). Visual speech information is provided by movements in the facial area: specifically, movements created by the face, the head and neck, and the lips. In order to be understood, speakers are assumed to strive to provide optimal acoustic and visual information to meet the demands of the target audience or the communicative situation (Burnham et al., 2002). Several studies (e.g., Burnham et al., 2000, 2001; Mixdorff & Charnvivit, 2004;

Mixdorff et al., 2005a, 2005b) have shown that visual speech information is related to the production of lexical tones. When speakers want to convey information about tone (the pitch contour for instance), facial cues (along with gestures) are a common visual resource they resort to alongside the acoustic information (Swerts & Krahmer, 2008; Zheng et al., 2018). Because our mouth, face, and head needs to move in a certain way to produce a given tone, the amplitude (range) and the length (duration) of the visible articulations change. For example, in Mandarin Chinese tones, there are clear differences in the duration of the vowels and the amplitude across tones: tone 3 usually has the longest vowel duration, while tone 4 tends to be the shortest; the amplitude for tone 3 is usually the lowest one, whereas tone 4 normally has the highest amplitude (Tseng, 1981). It makes much sense that these acoustic differences, for instance in the amplitude and the length of the articulation, have correlating visual characteristics (Chapter 2). Physiological studies (e.g., Xu & Sun, 2002) suggest certain restrictions with respect to the coordination of the laryngeal and articulatory systems, which may lead to visual cues for tones (Mixdorff et al., 2005a). In addition, regarding prosodic features, a significant correlation has been found between the motion of the head and fundamental frequency during the production of speech (Yehia et al., 2002). For example, in the case of the Mandarin tone 3 (a low-dipping tone in terms of height and contour), the correlated direction of head/neck motion during tone production is usually signaled by a low-falling-rising head movement.

However, the extent to which auditory-visual information facilitates or improves tone identification compared to auditory-only information (i.e., the superiority of bimodal performance compared to unimodal performance) differs widely across individuals (Grant & Seitz, 1998). Furthermore, the benefit of visual/facial information for tone perception depends strongly on context, and in particular on the availability of a clear and reliable acoustic signal. In situations where such a signal is available, extra visual information may actually distract the perceivers instead of facilitating their tone perception, since they are reluctant to use the visual information when acoustic sources are available and reliable. For example, Burnham et al. (2001) have found that in an experiment using clean speech, Australian English speakers performed better in a task of identifying Cantonese words that differed only in tone in the auditory-only (AO) condition than in the auditory-visual (AV) condition (where they had access to lip and face movements).

In our study, we look into the effects of modality and musicianship on Mandarin tone perception. More specifically, we presented musicians and non-musicians with auditory-visual or auditory-only tone stimuli. Because of extensive musical training, musicians are particularly sensitive to the acoustic structure of sounds (i.e., frequency, duration, intensity and timbre parameters). This sensitivity has been shown to influence their perception of pitch contours

in spoken language (Schön et al., 2004), but the extent to which musicians are affected by the presence of (exaggerated) visual information during speech perception has remained largely unexplored. Besides, while they are obviously related, pitch perception is not the same as the identification of lexical tone. While musicians might benefit from the additional information just like non-musicians, this is not a given. Given their extensive training to analyze the acoustic signal, they might not be as inclined to use visual cues (compared to non-musicians). Thus, they may benefit less from the added visual information. Musicians may have developed the ability to focus on specific properties of sounds and this superior ability may in turn help them categorize the sounds and make the relevant decision (Besson et al., 2011). We hypothesize that musicians may still benefit from the added visual information for the Mandarin tone identification, but that this contribution is likely *smaller* than that for non-musicians. While performing tone identification, our participants are exposed to different stimuli to which they have to respond. Over the course of the task, we expect participant's performance to improve progressively. In order to investigate the learning process and to see whether the two participant groups differ with respect to their learning rate, for example, whether the musicians learn faster, or display superior performance from the beginning, we will look at the performance over time. In general, we assume that performance improves with training. Whether musicians outperform non-musicians from the beginning, or show a steeper learning curve, is an open question.

In sum, we investigate the effects of musical ability on Mandarin tone identification by tone-naïve listeners (speakers of Dutch), with a specific interest in how their performance is mediated by differences in modality. The Gold-MSI was used to measure the musical sophistication of each participant. We conducted a linear regression analysis to find out whether a specific musical ability/skill as measured by the subscales of the Gold-MSI is related to successful tone identification. Since the effects of our two independent variables might vary among tones, we subsequently assess the effects for each tone individually in our study.

3.2 Materials and methods

A 2 (musical ability) × 2 (modality) between-participant design was employed in this study. Two groups of participants (musicians and non-musicians) were randomly divided over two modality conditions (auditory-visual vs. auditory-only). Given the likelihood of learning effects, it was not possible to include modality as a within-participant factor. Accuracy, defined as the percentage correct identification of a tone based on its auditory realization, was the dependent variable.

3.2.1 Participants

There were 170 participants comprising two groups that differed in musical ability: 86 non-musicians (mean age 22, 62 females) were recruited from the Tilburg University participant pool; 84 musicians (mean age 22, 35 females) were recruited from the Fontys School of Fine and Performing Arts (located in Tilburg). Eighty-three percent of the participants were native speakers of Dutch, with the remaining participants reporting German, French, Greek, English, Portuguese, Spanish, Italian, Russian, Indonesian, Bengali and Arabic as their native language. None of them were native speakers of tone languages, and none had had formal training to learn a tone language. The musician group consisted of participants who had had eight or more years of intensive music training and practice up until 2017, while none of the non-musicians had received continuous musical training[8]. A self-reported musical sophistication questionnaire, the Gold-MSI, was used to assess the musical skills and behaviors of the participants.

3.2.2 Materials and stimuli

Gold-MSI

Individuals differ in their repertoire of musical behaviors and in the level of skill they display for particular musical behavior (Müllensiefen et al., 2014). The Gold-MSI is an attested self-assessment instrument that measures individual differences on multiple dimensions towards musical skills and behaviors. Thirty-eight items in total measure individual differences in musical sophistication. Among them, 31 items are rated on a seven-point scale (1= *completely disagree* and 7= *completely agree*); for the remaining 7 items, participants choose one answer from 7 options (the first option yields 1 point; the seventh option yields 7 points (e.g., I can play 0 / 1 / 2 / 3 / 4 / 5 / 6 or more musical instruments).

The Gold-MSI is a multi-faceted instrument that measures different aspects of musical sophistication. It has five sub-scales and one general score for the following facets: *active engagement* comprised of nine items covering a range of active musical engagement behaviors (e.g., " I spend a lot of my free time doing music-related activities"); *perceptual abilities* also with nine items, most of them related to musical listening skills (e.g., "I am able to judge whether someone is a good singer or not"); *musical training* combines seven items about the extent of musical training and practice (e.g., "I have had formal training in music theory for __ years"); *singing abilities* consists of seven items that reflected different skills and activities related to singing (e.g., "I am able to hit the right notes

[8]Some of the non-musicians periodically had had some musical education, for example in their middle school.

when I sing along with a recording"); *emotions* included six items describing active behaviors in response to music (e.g., "I sometimes choose music that can trigger shivers down my spine"); the *general musical sophistication* had 18 items which incorporated representative questions from all the five sub-scales.

The Gold-MSI is used in this study to measure the individual's musicality. The factor structure and internal reliability of the Gold-MSI have previously been tested with a German sample (Schaal et al., 2014), and validated for use with secondary school pupils in a large German sample of 11-19 years old (Fielder & Müllensiefen, 2015); it has also been used in a study with young and older Dutch adults (Vromans & Postma-Nilsenová, 2016).

Stimulus construction

We constructed a word list with 10 Mandarin monosyllables (e.g., *ma, ying* …, based on stimulus material from Francis et al., 2008 and from Chen & Massaro, 2008). Each of these syllables was chosen in such a way that the four tones would generate four different meanings, resulting in 40 (10 syllables × 4 tones) different existing words in Mandarin Chinese (See Appendix 1 for a complete list of the stimuli, previously used by Chapter 2).

Material Recording

Four (2F, 2M) native Mandarin Chinese speakers were instructed to produce the 40 words in two different scenarios in sequence: a natural mode ("pronounce these words as if you were talking to a Chinese speaker") and a teaching mode ("pronounce these words as if you were talking to someone who is not a Chinese speaker"), with the recording of the natural stimuli preceding the recording of the teaching style stimuli. In both conditions, there were no other instructions or constraints imposed on the way the stimuli should be produced. A 20-minute break was given to the speakers between the two recordings to counter fatigue. Speaking style as a factor is not reported on in the current paper.[9]

The images and sounds from the speakers were recorded by Eye-catcher (version 3.5.1) and Windows Movie Maker (2012). One of the advantages of the Eye-catcher system is that the camera is located behind the computer screen and thus records people "through" the screen, which is convenient for unobtrusively capturing the full-frontal images of speakers' faces, similar to what listeners see in a face-to-face setting.

[9]An analysis with speaking style as one factor has been reported in Chapter 2. This paper focused on the effects of musicality and modality on tone identification for tone-naïve participants (experimental stimuli were the same as the previous paper, but there was a different group of participants). Including speaking style as a factor in the analyses did not meaningfully alter the effects of the other independent variables.

In total, two sets of 160 video stimuli (10 syllables × 4 tones × 4 speakers) were produced in teaching and in natural modes. These video clips were segmented into individual tokens, with each token containing one stimulus. We used Format Factory (version 3.9.5) to extract the sound from each video to generate stimuli for the auditory-only conditions. This resulted in four types of experimental stimuli: video + teaching (VT); video + natural (VN); audio + teaching (AT); audio + natural (AN), with each set containing 160 testing stimuli. Therefore, the auditory-visual (video) conditions include VT and VN, and the auditory-only (Audio) conditions are AT + AN.

3.2.3 Procedure

This study was carried out in accordance with the recommendations of the Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences, Tilburg University. The protocol was approved by the Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences, Tilburg University. All participants gave written informed consent in accordance with the Declaration of Helsinki.

The task of the participants was to (learn to) identify the tones they perceived from auditory-visual or auditory-only stimuli. We used E-prime (Version 2.0; Zuccolotto et al., 2012) to set up and run the experiment. Upon their arrival, participants signed an informed consent form that contained information about the nature of the experiment and their voluntary participation in it, agreeing for the data to be used for scientific research. They then filled out the Gold-MSI questionnaire, assessing their musical background. Next, they received a brief instruction about Mandarin Chinese tones. This instruction was displayed on the screen (see Figure 3.1 for a screenshot): "there are four tones in Mandarin Chinese: the first tone is a high-level tone, symbolized as " ˉ ", the second tone is a mid-rising tone, symbolized as " ˊ ", the third tone is a low-dipping tone, symbolized as " ˇ ", and the fourth tone is a high-falling tone, symbolized as " ˋ ".

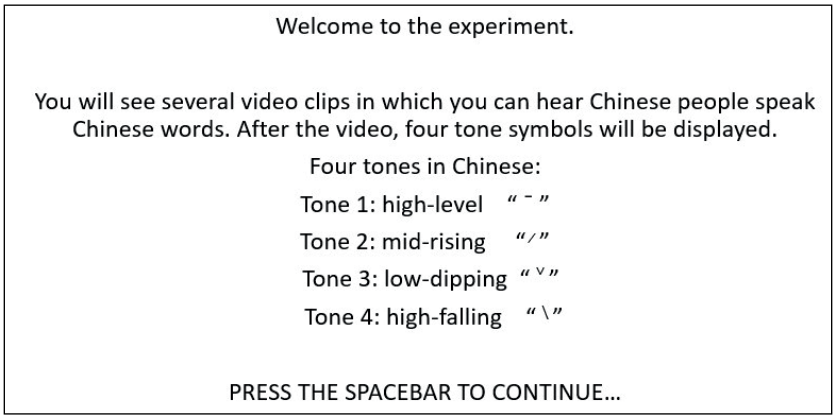


Figure 3.1. Screenshot of a brief introduction of Mandarin Chinese tones (in auditory-visual conditions)

The introduction was followed by exposure to three practice trials, either auditory-only or auditory-visual, depending on the condition they were randomly assigned to. After those, the experiment leader checked whether they had fully understood the concept of tones (in particular the symbols) and the task was clear. The main experiment consisted of 160 testing stimuli (video/audio), which were presented in an individually randomized order (operated by E-Prime). Participants received feedback in both the practice and testing trials in the form of a "good job" or "incorrect" message on the screen after their response. If no response was recorded within 10 seconds after the end of the stimulus, "no response" was shown. This registered as a missing response.

Participants were seated in a sound-attenuated room, wearing headsets, directly in front of the PC running the experiment. All stimuli were presented at a comfortable hearing level. They were told to press the designated keys with the corresponding tone symbols (" ¯ ", " ˊ ", " ˇ ", " ˋ ", see Figure 3.2) on them as accurately and as quickly as possible after they made their decisions. Their responses (and reaction times) were recorded automatically by E-prime.

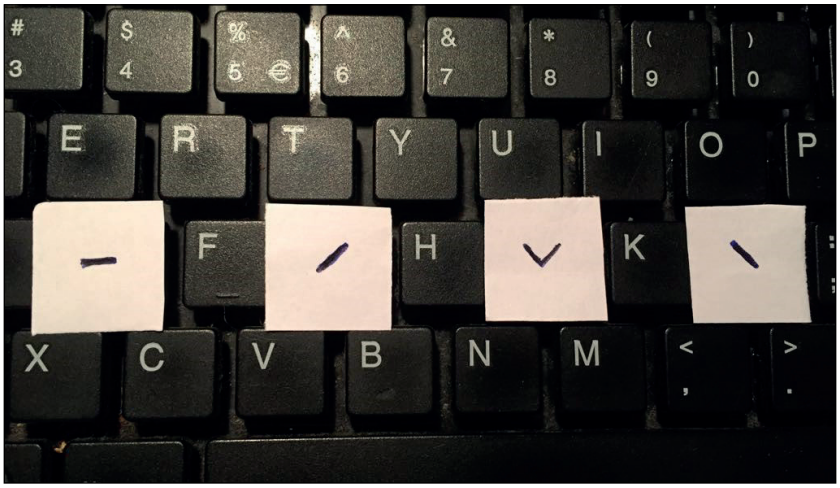


Figure 3.2. Picture of the designated keys with tone symbols on them

3.3 Results

3.3.1 Overall tone perception

In order to examine to what extent modality (auditory-visual vs. auditory-only) and musical ability (musicians vs. non-musicians) affect the perception of Mandarin Chinese tones, a mixed ANOVA was carried out with modality and musical ability as between-subject factors, and speaker and tone as within-subject factors. The percentage of correct responses (accuracy) was analyzed as the dependent variable.

Figure 3.3 depicts the performance of musicians and non-musicians in the two experimental conditions. Overall, participants were able to identify Mandarin tones well above chance (25%) (a histogram of each participant's accuracy in Appendix 2 shows that only about 1.8%, i.e., 3 individuals, of the participants score below chance over all tones), and the musician group outperformed the non-musician group in both experimental conditions, as indicated by a higher percentage of correct responses (M = 75%, SE = 0.02 vs. M = 48%, SE = 0.02). The difference in percentage correct between musicians and non-musicians was statistically significant (F (1, 166) = 150, p < .001, η_p^2 = .48), which was in line with our hypothesis that musical ability positively affects the ability to identify Mandarin tones.

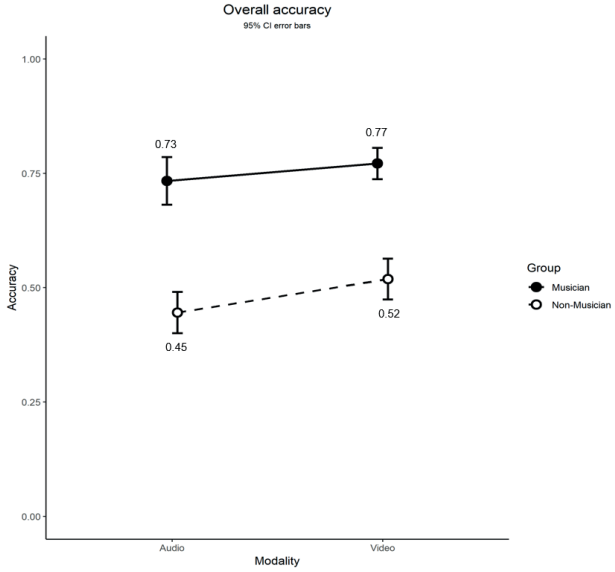


Figure 3.3. Average accuracy in percentage correct of Mandarin tone identification as a function of musical background and modality. Video represents auditory-visual conditions and Audio represents auditory-only conditions. Error bars represent standard errors.

The statistical analyses further showed that the auditory-visual condition (M = 65%, SE = 0.02) yielded significantly higher accuracy scores than the auditory-only condition (M = 59%, SE = 0.02); F (1, 166) = 6.39, p = .012, η_p^2 = .037. These results are in line with the hypothesis that the availability of visual cues along with auditory information is useful for people who have no previous knowledge of Mandarin Chinese tones when they need to learn to identify these tones. For musicians, seeing the speaker (video condition) helped them to identify more tones correctly as compared to only listening to the speaker (audio condition): 77% vs. 73%, t (13438) = 5.12, p < .001. For non-musicians, the effect of visual information was even greater: 52% in auditory-visual condition and 45% in auditory-only condition, respectively; t (13758) = 8.60, p < .001. Notably, there was no significant interaction between musicality and modality: F (1, 166) = 0.66, p = .42, η_p^2 = .004, which indicates that the effects of musicality or modality on tone perception are not dependent on each other. Overall, the lack of two-way interaction indicates that both groups identified tones more accurately in the video condition than in the audio condition.

Musicians indeed outperformed non-musicians in terms of average accuracy in Mandarin tone identification. However, as we mentioned in the first section of this study, the learning patterns over time for both participant groups are of interest in this regard as well. We included stimulus number (1-160) as a predictor in the regression model. In the regression with stimulus number as a predictor (both separately and together with musicianship and modality) and the average accuracy as the dependent variable, stimulus number significantly predicts performance: F (1, 638) = 32.82, p < .001, with an R^2 of .049 and F (3, 636) = 829, p < .001, with an R^2 of .796, respectively. We also did a regression analysis with the interaction between stimulus number and musicianship (to see whether the learning rate of musicians and non-musicians differed), but that effect was small and non-significant (p = .057). The scatter plots in Figure 3.4 show the learning curves for musicians and non-musicians in audio (auditory-only) and video (auditory-visual) conditions. In general, both groups of participants showed improvement in performance over time, which shows both groups are learning. Although musicians performed better than non-musicians across the board (they had a higher accuracy), they did not learn faster than non-musicians.

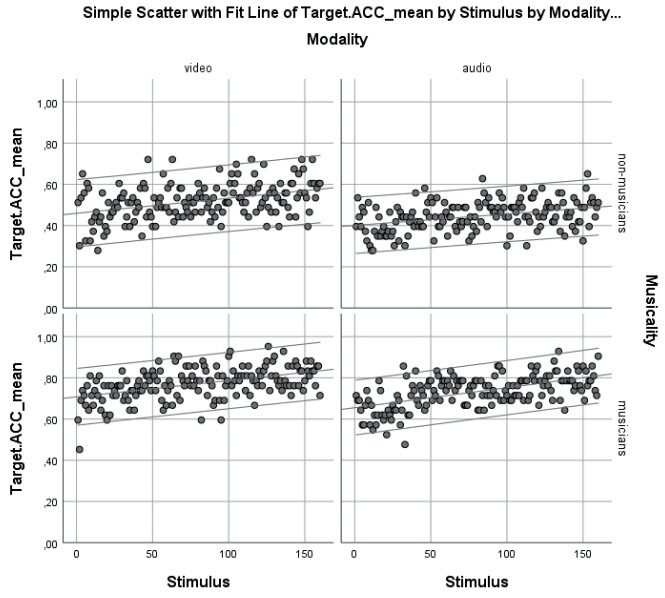


Figure 3.4. Learning curves for musicians and non-musicians in Audio (auditory-only) and Video (auditory-visual) conditions

3.3.2 Individual tone perception

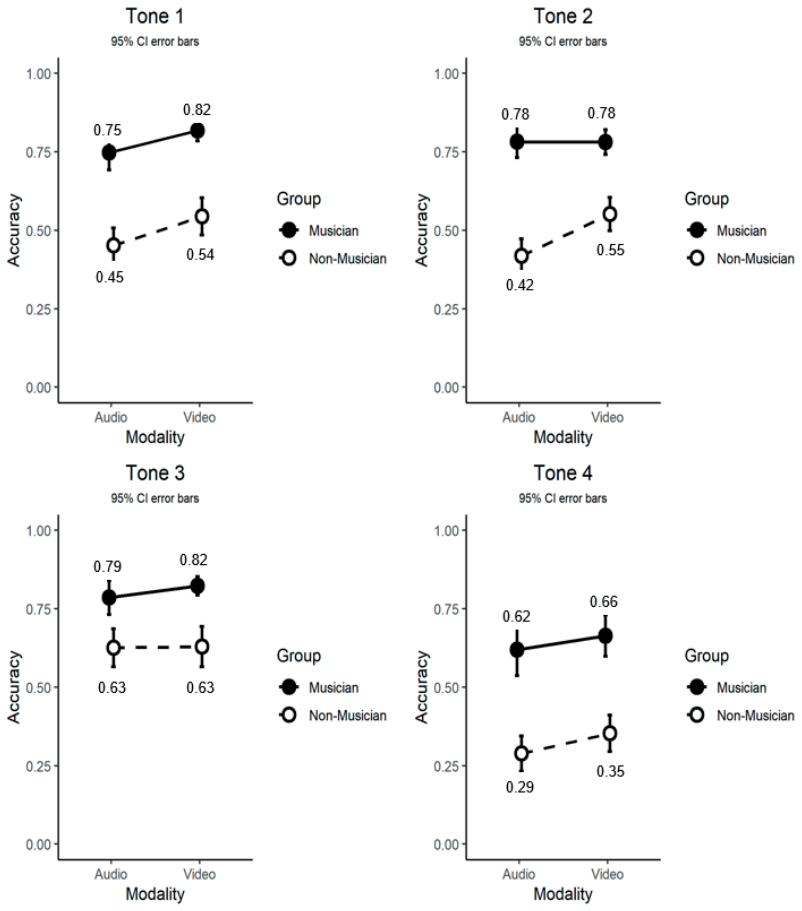


Figure 3.5. Average accuracy in percentage correct of Mandarin tone identification as a function of modality, musicality and tone. Video represents auditory-visual conditions and Audio represents auditory-only conditions. Error bars represent standard errors.

Figure 3.5 shows the identification performance in terms of accuracy of musicians and non-musicians in the two experimental conditions for each of the four Mandarin tones. Musicians again performed better than non-musicians for all four Mandarin tones. While musicality continued to play a significant role in each tone's identification: F (3, 498) = 10.76, p < .001, η_p^2 = .061, modality did not affect the accuracy of identifying each tone: F (3, 498) = 1.79, p = .149, η_p^2 = .011. For both participant types (and for all combinations of modality) tone had a strong effect on accuracy (F (3, 498) = 102, p < .001, η_p^2 = .38). Furthermore, there was a significant two-way interaction between tone and musicality and a significant three-way interaction between tone, musicality, and modality (F (3, 498) = 3.33, p = .022, η_p^2 = .02).

Given the significant three-way interaction we conducted four separate (one for each tone) two-by-two analyses of variance with musicality and modality as independent variables and accuracy as the dependent variable. For tone 1, this resulted in a main effect of modality (F (1, 169) = 9.99, p = .002), with accuracy being higher in the video condition, a main effect of musicality (F (1, 169) = 122.18, p < .001), with accuracy being higher for musicians, but no significant interaction between the two main effects (F (1, 169) = 0.19, p = .66). For tone 2, we found a main effect of modality (F (1, 169) = 7.15, p = .008), with accuracy being higher in the video condition, a main effect of musicality (F (1, 169) = 144.82, p < .001), with musicians being more accurate than non-musicians, but this effect was qualified by a significant interaction between musicality and modality (F (1, 169) = 7.45, p = .007). This significant interaction was further analyzed in two independent t-tests contrasting the auditory-visual and the auditory-only modality separately for each participant group. For musicians, this analysis showed no significant difference between the two modalities; t (82) = -0.04, p = .97. Non-musicians however, were significantly more accurate in the video condition; t (84) = 3.55, p = .001. For tone 3, there was no main effect of modality (F (1, 169) = 0.576, p = .450), but the main effect of musicality (F (1, 169) = 43.46, p < = .001) was present, again with musicians being more accurate. As with tone 1, there was no significant interaction between musicality and modality (F (1, 169) = 0.40, p = .556). Similarly, for tone 4, there was no main effect of modality (F (1, 169) = 2.79, p = .097), but musicians scored better than non-musicians (F (1, 169) = 95.79, p < .001), with accuracy being higher for musicians, but no significant interaction between the two main effects (F (1, 169) = 0.092, p = .762).

In sum, the main effect of musicality holds for each tone: musicians performed better than non-musicians with all tones, the significant two-way interaction points towards differences of degree. In contrast, the effect of modality is present only for tone 1, where performance in the video condition is consistently superior. There is a main effect for tone 2, but this is driven by the significant two-way interaction where musicians are unaffected by modality, but

non-musicians do perform better in the auditory-visual condition compared to the auditory-only condition.

The tonal confusion matrices below (Table 3.1a for musicians and Table 3.1b for non-musicians) give more insight in the way our participants perceived individual Mandarin tones. The data show that, regardless of the fact that musicians performed much better than non-musicians on the identification of each tone, both groups of participants can identify the tones above chance (25%). For both groups, the low-dipping tone 3 was the easiest to recognize (80.4% and 62.7% for musicians and non-musicians), while the high-falling tone 4 was the most difficult to identify (64.2% and 32.1% respectively). In general, when hearing tone 4, participants often confused it with the high-level tone 1, though the confusions were not necessarily symmetrical: tone 1 was mostly confused with mid-rising tone 2, rather than with tone 4. The participants seemed to be able to identify equally well tone 1 and tone 2. These confusions are similar to the ones we find in our previous study (Chapter 2).

		Responded tone			
		1	2	3	4
Presented tone	1	**78.3**	10.8	3.0	7.9
	2	3.6	**78.2**	14.9	3.4
	3	0.9	12.2	**80.4**	6.5
	4	23.1	9.0	3.6	**64.2**

Table 3.1a. Confusion matrix for tone (percentage correct) in musicians

		Responded tone			
		1	2	3	4
Presented tone	1	**49.8**	26.3	8.9	15.0
	2	16.8	**48.5**	20.7	14.1
	3	5.2	15.1	**62.7**	17.0
	4	34.7	23.0	10.3	**32.1**

Table 3.1b. Confusion matrix for tone (percentage correct) in non-musicians

3.3.3 A more fine-grained look at musicality

In the above analyses, we grouped the participants according to their affiliated institutions: musicians (participants from the Fontys School of Fine and Performing Arts) and non-musicians (participants from Tilburg University). However, while the sample of musicians was clearly more musical than the non-musicians, we did not quantify the extent of the difference, neither for musicality as a whole or for different aspects of musicality, nor did we take into account the possibility that, at least in some areas, there might be university students with considerable musical experience. To get a better handle on the musical abilities of both the musicians and non-musicians in our study, we employed the Gold-SMI questionnaire. With the five dimensions/sub-scales included in the questionnaire (active engagement, perceptual abilities, musical training, singing abilities, and emotions) as dependent variables, and group membership (musicians and non-musicians) as the independent variable, we conducted a multivariate analysis of variance (MANOVA) in order to get a more detailed picture of the differences and similarities in musicality between these two groups. Table 3.2 contains a summary of the outputs for the five dependent variables for musicians and non-musicians. On average, musicians attained a higher value when compared to non-musicians in each category.

	Non-musicians		Musicians			
Variable	Score	*SD*	Score	*SD*	F (1, 168)	η_p^2
Active Engagement	3.95	1.06	5.34	0.63	106.17**	.39
Perceptual Abilities	4.88	0.87	5.97	0.55	94.86**	.36
Musical Training	2.70	1.38	5.64	0.57	325.29**	.66
Singing Abilities	3.80	1.13	5.46	0.77	125.49**	.43
Emotions	5.02	0.86	5.86	0.63	52.03**	.24

Table 3.2. MANOVA results for non-musicians (N= 86) and musicians (N = 84). Scores range from 1-7. Note that the maximum score is 7; **p < .001.

Using Pillai's trace, there was a significant multivariate effect of group membership (musician vs. non-musician) on the five musical dimensions of the subject, V = 0.67, F (5, 164) = 67.6, p < .001, η_p^2 = .67, with musicians scoring higher than non-musicians on all five subscales. As shown in Table 3.2, significant univariate effects were also found for the five dimensions. The effect sizes (partial eta-squared) of the five subscales differ considerably, ranging from

.24 to .66, indicating marked differences in the importance of the subscales. The most prominent difference between musicians and non-musicians is in (reported) musical training ($\eta_p^2 = .66$).

3.3.4 Musicality and tone perception

In order to obtain a comprehensive view of the relation between musical experience of the participants and their tone perceptual ability, we first constructed a (Pearson) correlation matrix. Table 3.3 shows the correlations among the five sub-scales (active engagement, perceptual abilities, musical training, singing abilities, and emotions) and the accuracy of the participant's tone identification. The results indicated that there is a significant positive association between all sub-scales and the performance (accuracy).

		1	2	3	4	5	6
1	Active Engagement						
2	Perceptual Abilities	.686***					
3	Musical Training	.727***	.716***				
4	Singing Abilities	.630***	.807***	.717***			
5	Emotion	.700***	.678***	.561***	.560***		
6	Accuracy	.484***	.547***	.653***	.525***	.441***	

Table 3.3. Table of Correlations for Gold-MSI variables and accuracy of tone perception

We also conducted two linear regression analyses to see if the more fine-grained Gold-MSI scales predict anything above the binary classification between musicians and non-musicians. Specifically, we compared a linear regression with groups (musicians vs. non-musicians) as the predictor (Model 1) with a regression that also includes the five sub-scales of Gold-MSI as predictors (Model 2), using the overall accuracy as outcome variable. Table 3.4 contains the summary for the two models. The data in the table show that adding Gold-MSI predictors significantly improves the model ($R^2 = .47$ in Model 1 and $R^2 = .51$ in Model 2; $F_{change} = 2.99$, $p = .013$); and that musical training is the only Gold-MSI variable that predicts additional variance in identification accuracy: $b = 0.37$, $\beta = 0.24$, $t\,(163) = 2.08$, $p = .039$, although weakly so. None of the other predictors were significantly related to accuracy.

Model		*B*	*SE*	*β*	*t*	*p*
1	Groups	27.02	2.24	.68	12.09	.00
	$R^2 = .47$					
	$F_{change} = 146.09$					< .001
2	Groups	17.84	3.80	.45	4.70	.00
	Active Engagement	-.21	.19	-.11	-1.15	.25
	Perceptual Abilities	.41	.26	.17	1.55	.12
	Musical Training	.37	.18	.24	2.08	.04
	Singing Abilities	-.12	.22	-.05	-.53	.60
	Emotion	.31	.32	.08	.97	.33
	$R^2 = .51$, $R_{change}^2 = .05$					
	$F_{change} = 2.99$					.013

Table 3.4. Multiple linear regressions for accuracy of tone identification in Model 1 (Groups as the predictor) and Model 2 (Groups + five sub-scales as the predictors)

From the analysis above, the amount of musical training received emerges as the only predictor of accuracy of tone perception. To investigate whether the other predictors add anything to the effects of training considering both groups simultaneously, we repeated our regression analysis comparing a model with musical training as the predictor (Model 3) and a model with musical training plus the other four sub-scales of Gold-MSI as predictors (Model 4). As before, the data in Table 3.5 show that the other four predictors did not significantly improve the accuracy of the tone perception ($R^2 = .43$ in Model 3 and $R^2 = .44$ in Model 4; $F_{change} = 1.29$, $p = .28$).

Model		*B*	*SE*	*B*	*t*	*p*
3	Musical Training	1.02	.09	.65	11.17	.00
	$R^2 = .43$					
	$F_{change} = 124.81$					< .001
4	Musical Training	.87	.15	.56	5.62	.00
	Active Engagement	-.17	.20	-.09	-.87	.38
	Perceptual Abilities	.29	.28	.12	1.05	.29
	Singing Abilities	.07	.23	.03	.32	.75
	Emotion	.35	.34	.09	1.02	.31
	$R^2 = .44$, $R_{change}^2 = .02$					
	$F_{change} = 1.29$					.28

Table 3.5. Multiple linear regressions for accuracy of tone identification in Model 3 (musical training as the predictor) and Model 4 (musical training + four other sub-scales as the predictors)

The above analyses provide a general picture of the relationships between musicality (and the five subscales) and the accuracy of tone perception. As a final step, we zoomed in on the individual tones to see whether these particular factors predicted the perception of specific tones. Multiple regressions were conducted for musicians and non-musicians combined for each of the individual tones. The results showed that the musicality significantly predicted accuracy for each of the individual tones, and out of the five individual factors, musical training was the only constant factor of predicting the accuracy for each individual tone, while the other factors had no consistent effect on accuracy.

3.4 Discussion

We set out to investigate two factors that influence Mandarin tone perception in tone-naïve listeners: the musicality of the participants (comparing musicians and non-musicians) and the stimulus modality (comparing audio-visual and auditory-only stimuli). The findings of the study were:

(1) All participants were able to identify Mandarin tones well above chance level;
(2) Musicians outperformed non-musicians in both auditory-visual and auditory-only presenting conditions;
(3) The amount of musical training is the only factor that relates to successful tone identification;
(4) The auditory-visual condition yielded significantly better results than the auditory-only condition;
(5) The effect of musicality and modality on tone identification varies among individual tones.

We will discuss these findings one by one.

In line with previous studies, we replicate the finding that musicians are at an advantage compared to non-musicians when learning to identify lexical tones in Mandarin Chinese for non-native listeners (Alexander et al., 2005; Delogu et al., 2006, 2010; Gottfried & Riester, 2000; Gottfried et al., 2004; Lee & Hung, 2008). Based on our findings, we would argue that the length of musical training led (musicians) listeners to a better performance in Mandarin Chinese tone identification: listeners with more musical training showed considerably greater accuracy in their identification (75% vs. 48%). Importantly, although the musicians in our study performed well in the identification task (79% at the highest for the dipping tone 3), they did not achieve native-like performance (as reported in Chapter 2), and the learning patterns tell us that musicians did not learn faster than non-musicians. Musicians showed their superior performance at the beginning of the task. Interestingly, the increase in performance follows a linear path for both musicians and non-musicians, and does not seem to plateau, indicating that more exposure leads to better performance, and potentially (in the case of a longer learning period) may lead to still higher final accuracy scores.

Although musical training has been identified as the only factor that predicts tone identification, it is not a foregone conclusion that the other aspects of musicality do not affect the learning of Mandarin tones. Because our study uses natural groups of musicians and non-musicians, musical training is confounded with group membership. Importantly, if we analyze both groups separately, there is no relationship between musical training and

tone identification performance among the musicians and non-musicians. The absence of the relationship between musical training and tone identification in musicians might be due to a lack of variation in training among musicians (a restriction of range effect). Alternatively, since the Gold-SMI is originally intended for using in the general population, it may not be able to capture the more subtle differences among musicians in as much detail as is required to differentiate among musicians. In addition, the parts of the Gold-MSI we used all relied on self-report, which might not be able to capture important differences in factors such as perceptual abilities. However, and importantly, Müllensiefen et al. (2014) reported high correlations (ranging from .30 to .51) for the relation between self-report and objective listening performance (see page 9, for the AMMA listening test) and similarly in an online listening test (correlations ranging from .11 to .52). Future studies could include behavioral tests (also present in the Gold-MSI) to be able to better characterize the differences in musical skills and relate them to tone identification.

Nevertheless, our findings point to the interesting possibility of aiding language learning by providing learners additional musical training. Since musical training is the only consistent predictor for performance on the tone identification task, and training is something that potentially anyone can do – it is not a talent or innate ability – our results are promising for educational purposes. For example, second/foreign language learners could get some musical training to facilitate their language learning; schools can enrich students' curriculum with musical lessons; teachers may consider blending musical training into their language materials.

With respect to modality, tone-naïve listeners were able to identify tones better when they saw *and* heard the speakers compared to when they only heard them. This supports the hypothesis that visual information plays a facilitating role in learning to identify Mandarin tones for tone-naïve listeners, although the effect was not that large, with participants' accuracy increasing by 6% in the auditory-visual condition. Rather than distracting the listeners (as suggested by Burnham et al., 2001), the presence of facial expressions appears to facilitate Mandarin tone perception in clean speech. Both participant groups benefited from visual information, but numerically the non-musicians did so more than the musicians. This could be because musicians are trained to be particularly sensitive towards acoustic information, and they are already so good at identifying tones that the additional contribution of visual information is limited. This explanation is in line with our earlier assumption that musicians would benefit less from the added visual information compared to non-musicians. In our data, the modality effect is restricted to tones 1 and 2 (for non-musicians). This may be related to the intrinsic properties of individual tones, as tones differ in how easy they are to identify, and in the amount of auditory contour information they provide. For example, the auditory contour

of tone 1 (high-level tone) and 2 (mid-rising tone) is much less pronounced than that of tone 3 (low-dipping tone) and 4 (high-falling tone). As the auditory information is often mimicked in facial expressions (see for example, Swerts & Krahmer, 2008), there is simply less auditory information to transfer to the visual domain. Regardless of their specific contour, the tones differ in their overall difficulty, with tone 3 being the easiest and tone 4 being the hardest (as shown in the confusion matrices). It might be that our non-musicians ignored visual information in tone 3 because auditory information was sufficient for them, and, also ignored visual information for tone 4, because combining the auditory and visual information is too challenging. In contrast, tone 1 and 2 present the sweet spot where perceivers are able to take both auditory and visual information into account.

Our findings imply that (non-native) listeners learning Mandarin tones might benefit from pointing out the information that visual cues can contribute. Although we do not really know yet what the exact visual cues are, or in other words, what the listeners should look at, our finding is a good starting point for further exploration. For instance, in a teaching context, teachers should consider using their facial expressions while talking to students/learners, while learners/students could be trained to consciously attend to visual information. Similarly, in online learning environments, using video is likely to be more effective than using audio-only material.

The extent to which the listeners can benefit from visual cues also depends on individual speaker characteristics. There are substantial differences in the degree to which the speakers' faces exhibit relevant characteristics (Chapter 2; Bradlow & Bent, 2002; Gagné et al., 1994). Most previous studies have concluded that female speakers in general are better than male speakers are at displaying salient articulation, such as expanding their overall vowel space and increasing their F0 mean (Cox et al., 1987; Ferguson, 2004; Ferguson & Kewley-Port, 2007; Kricos & Lesner, 1982). However, due to the limited number of speakers, our study does not allow us to draw conclusions about gender differences, or speaker differences for that matter. Nevertheless, further research should take into account the variations between speakers' realizations of visual information.

Crucially, individual tones are important contributors to the observed differences in tone identification. In other words, it is more important which tone the listeners hear than the modality in which it is presented. The low-dipping tone 3 is the easiest one to identify, while all listeners had more difficulty identifying the high-falling tone 4, and this holds for both musicians and non-musicians in both experimental conditions (auditory-visual and auditory-only). This is possibly due to their specific temporal characteristics – tone 3 has the longest duration and two intensity peaks, while tone 4 has the shortest duration, and only one intensity peak. Our findings with respect to

the accuracy differences between tones differ somewhat from previous studies: while tone 3 has indeed consistently been found the easiest to identify, due to the longest vowel duration, tone 4 has not always been found to be the most difficult one to recognize (Blicher et al., 1990; Fu & Zeng, 2000; Mixdorff et al., 2005b). Nevertheless, it is clear that individual tones differ in learnability, which, too, is relevant when considering teaching Mandarin tones (for example when designing a curriculum).

3.5 Conclusion

In sum, the present study contributes to the literature on the relationship between musicality and tone identification, and the roles played by auditory and visual speech information. The results showed that musical training in particular facilitates Mandarin tone perception. Furthermore, learning Mandarin tones can be facilitated by being aware of the information provided by both the auditory and the visual modality. Finally, it is clear that the individual tones differ in how easy they are to identify. We aim to investigate the contributions of these factors in future work and hope that our findings will benefit second language learners of Mandarin and will inspire further research on Mandarin tone learning.

Chapter 4

Chapter 4

Relative Contribution of Auditory and Visual Information to Mandarin Chinese Tone Identification by Native and Tone-naïve Listeners

Abstract. Speech perception is a multisensory process: what we hear can be affected by what we see. For instance, the McGurk effect occurs when auditory speech is presented in synchrony with discrepant visual information. A large number of studies have targeted the McGurk effect at the segmental level of speech (mainly consonant perception), which tends to be visually salient (lip-reading based), while the present study aims to extend the existing body of literature to the suprasegmental level, that is, investigating a McGurk effect for the identification of tones in Mandarin Chinese. Previous studies have shown that visual information does play a role in Chinese tone perception, and that the different tones correlate with variable movements of the head and neck. We constructed various tone combinations of congruent and incongruent auditory-visual materials (10 syllables with 16 tone combinations each) and presented them to native speakers of Mandarin Chinese and speakers of tone-naïve languages. In line with our previous work, we found that tone identification varies with individual tones, with tone 3 (the low-dipping tone) being the easiest one to identify, whereas tone 4 (the high-falling tone) was the most difficult one. We found that both groups of participants mainly relied on auditory input (instead of visual input), and that the auditory reliance for Chinese subjects was even stronger. The results did not show evidence for auditory-visual integration among native participants, while visual information is helpful for tone-naïve participants. However, even for this group, visual information only marginally increases the accuracy in the tone identification task, and this increase depends on the tone in question.*

*This chapter is based on: Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2020). Relative Contribution of Auditory and Visual Information to Mandarin Chinese Tone Identification by Native and Tone-naïve Listeners. *Language and speech*, 63(4), 856-876. https://doi.org/10.1177/0023830919889995

4.1 Introduction

Speech perception is more than just an auditory event: it is a multisensory/multimodal process (Campbell, Dodd, & Burnham, 1998; Massaro, 1998). What we *hear* can be affected by what we *see*. For instance, seeing the face of the speaker normally helps the listener perceive speech better (Bailly et al., 2012; Hirata & Kelly, 2010; Sumby & Pollack, 1954), especially in noisy environments (e.g., Burnham et al., 2001; Mixdorff et al., 2005b). Similarly, seeing the face of a speaker also aids hearing impaired listeners in decoding the auditory speech signal (Desai et al., 2008; Smith & Burnham, 2012).

One of the possible reasons why visual information benefits human speech perception is that it provides complementary information about the place of articulation, which is sometimes difficult to deduce from auditory information alone (Binnie, Montgomery, & Jackson, 1974). For example, the unvoiced consonants /p/ (a bilabial) and /k/ (a velar), the voiced consonant pair /b/ and /d/ (a bilabial and alveolar, respectively), and the nasal /m/ (a bilabial) and the nasal alveolar /n/ (Massaro & Stork, 1998) are minimal pairs that are easy to confuse based on auditory information alone (Potamianos, Neti, Gravier, Garg, & Senior, 2003). Facial information, such as the shape of the lips, the position of the jaw, and the motion of the cheeks helps listeners disambiguate between such confusable minimal pairs (Jiang, Alwan, Keating, Auer, & Bernstein, 2002).

While congruent visual information during articulation generally improves speech perception (Cutler & Chen, 1997; Ye & Connine, 1999), discrepant visual information can alter speech perception, which has been exemplified in the now classic McGurk effect (McGurk & MacDonald, 1976). McGurk and MacDonald demonstrated how visual information about the place of articulation (lip movements) can modify phonetic perception: observers perceived an auditory [ba] paired with a visual [ga] as "da". Access to visual information about the source of speech can thus have clear effects on speech perception. This perceptual fusion between auditory and visual information is caused by the fact that the human visual system is highly sensitive to the distinction between labials (/b/ and /m/, for instance) and non-labials (such as /d/ and /n/) (Sekiyama, 1997). In other words, with the McGurk effect, visual information that is discrepant (in terms of place of articulation—lip movements) with the auditory signal misleads and biases perceptual judgment, whereas it normally helps auditory perception in the natural auditory-visual congruent situation.

Since McGurk and MacDonald (1976) first reported this fusion effect between auditory and visual information, a number of studies have been carried out across languages to investigate the nature of the effect with various combinations of auditory and visual syllables. The McGurk effect has been found in native speakers of various languages: for instance, English (McGurk

& MacDonald, 1976), German and Spanish (Fuster-Duran, 1996), Dutch and Cantonese (de Gelder, Bertelson, Vroomen, & Chen, 1995), Italian (Bovo, Ciorba, Prosser, & Martini, 2009), Thai (Burnham & Dodd, 1996), Japanese (Sekiyama & Tohkura, 1991) and Chinese (Sekiyama, 1997). The majority of the studies tested one single language with native subjects; other studies tested one language with non-native and native subjects, for example, Austrian German with Hungarian subjects (Grassegger 1995) and a series of cross-culture/intercultural studies on the McGurk effect that tested two languages with native and non-native subjects was also carried out to examine the inter-language differences in terms of the magnitude of the McGurk effect (Burnham & Dodd, 2018; Hayashi & Sekiyama, 1998; Sekiyama, 1997; Sekiyama &Tohkura, 1991).

The McGurk effect has been established as a language- and culture-dependent phenomenon: there is a robust McGurk effect in English-speaking languages/cultures, while it is relatively weak in Asian languages/cultures. Comparing native speakers of Japanese with native speakers of American English, Sekiyama (1994) reported that the English subjects showed a larger McGurk effect than the Japanese subjects. Subsequently, Sekiyama (1997) found that the native speakers of Japanese showed a larger McGurk effect than Mandarin Chinese speakers. In line with these results, Burnham and Lau (1998) found a larger McGurk effect in English speakers as compared to Cantonese speakers.

Sekiyama (1997) proposed two major factors to explain why there are inter-language differences in the McGurk effect (weaker in Asian languages, stronger in English-speaking cultures). One is a cultural factor, which has been developed as the face-avoidance hypothesis. In some Asian cultures, like the Japanese and Chinese, as a social rule, it is discouraged to directly look at the speakers, which might suppress access to the information needed to integrate the visual stimuli with auditory information, even in a face-to-face communicative setting. The other factor is based on a linguistic characteristic of many Asian languages, and is known as the tone hypothesis. Tonal languages (such as Mandarin) and semi-tonal languages (such as Japanese) have fewer phonemes (consonants, vowels and syllables) and a simpler syllabic and phonological structure (in Japanese, at least) compared to English. Because of this, the lip-read information may be used less in speech processing (Sekiyama & Burnham, 2008). Therefore, the more tonal the language, the greater the reliance on auditory information, and thus a less strong McGurk effect (Burnham & Lau, 1998; Magnotti et al., 2015).

In order to test the tone hypothesis, Burnham and Lau (1998) explored the effect of tonal information on auditory reliance in the McGurk effect, by presenting both tonal (Cantonese) and non-tonal (English) language speakers with McGurk stimuli ([ba] [ga]) in which the tone on syllables either varied or remained constant (pronounced by Cantonese and Thai speakers) across trials. They found that Cantonese subjects relied more on auditory information alone

than (Australian) English subjects did; this reliance on auditory information was stronger in the condition with tone variation compared to stimuli where tone was kept constant.

Crucially, tone languages, such as Mandarin Chinese, do not only rely on phonological distinctions between vowels and consonants, but additionally use tones to distinguish word meanings. This is different from most European languages, which almost exclusively rely on phonological distinctions at the segmental level. For instance, if the Mandarin Chinese syllable /ma/ is produced with a rising tone, it means "hemp", whereas it means "scold" when produced with a falling tone. Pitch accent languages, such as Japanese, also have some tonal properties (high and low pitch), but to a much smaller extent than Mandarin Chinese. Scholars such as Sekiyama (1997) and Magnotti et al. (2015) have explored the McGurk effect in native speakers of Mandarin Chinese (as described above), although in all cases they targeted the McGurk effect at the segmental level of speech (mainly consonant perception). Consonant perception is fairly susceptive to visual information, because place of articulation is a major determinant (i.e., lip-read), and that is relatively more visually salient, while the present study extends the auditory-visual integration to the suprasegmental level, that is, the four Mandarin Chinese tones.

Previous studies have shown that visual information plays a role in Chinese tone perception (e.g., Chen & Massaro, 2008; Chapter 2; Mixdorff et al., 2005; Reid et al., 2015; Shaw, Chen, Proctor, Derrick, & Dakhoul, 2014), although the effects of visual information are subtle. For example, based on visual information only, native speakers of Cantonese can still distinguish Cantonese tones significantly above chance under certain conditions (Burnham, Ciocca, & Stokes, 2001). Similarly, Chen and Massaro (2008) asked Mandarin perceivers to identify Mandarin Chinese tones in the visual-only condition, and they found that the performance of native speakers is statistically significant above chance. The fact that visual information does provide relevant cues for tone identification points to the potential of multisensory integration at the tone level, possibly leading to a McGurk effect. Although it is unclear what the exact visual cues are for tone identification, there is some evidence for the existence of visual cues for individual Mandarin tones. Specifically, tone identification has been found to mainly depend on the (intensity of the) movements of the mouth, head/chin, and neck: specifically, there is little to no activity for tone 1, some activity for tone 2 and tone 4 (although very brief for tone 4), with tone 3 having the most activity, namely a dipping head/chin. Duration (time) differences between the tones may be caused by variation in the movements of the mouth, as more complex movements would require more time to be realized (Chen & Massaro, 2008). Similarly, Vatikiotis-Bateson and Yehia (1996) and Yehia et al. (2002) found strong correlations between head movements and F0. Such visual cues that relate to more general movements of the head have previously

also been reported to function as correlates of larger-scale prosodic structures in other languages, for example, quick movements of the head that co-occur with pitch accents (Krahmer & Swerts, 2007). Whether there is auditory-visual integration in Mandarin Chinese in the form of a tonal McGurk effect is one of the two main research questions of this study. To answer this question, we constructed various combinations of congruent and incongruent auditory and visual tone stimuli and presented them to test Chinese participants.

The other main question is whether visual information affects tone perception for non-native speakers differently. More specifically, we investigate the relative contribution of auditory and visual information in Mandarin Chinese tone identification in tone-naïve speakers. Sekiyama argued in her study (1994) that Japanese listeners as native speakers are sensitive to the discrepancy and incompatibility between the auditory and visual information in the cross-dubbed material, and they therefore tend to separate the conflicting visual information from the auditory information, when audition provides sufficient information (i.e., in a noise-free speech condition). The American participants in the study, on the other hand, showed a larger McGurk effect, because they tend to integrate the information when they perceive the stimuli as unintelligible as evidenced by the fact that the magnitude of the McGurk effect is the largest when American participants were presented with Japanese stimuli (leading to the so-called intelligibility hypothesis by Sekiyama in 1997). In addition, apart from a difference in the strength of the effect, the pattern of confusion (i.e., how the auditory percept is affected by visual cues) may also differ between groups of participants, given that the tones are phonologically relevant for only one of the compared languages.

In summary, we aim to answer two questions in this study: the first question is whether a McGurk effect can also be discerned at the tone level in native speakers of Mandarin Chinese. Secondly, we want to know how visual information affects tone perception for native speakers and non-native (tone-naïve) speakers. More specifically, we compare the relative contribution of auditory and visual information during Mandarin Chinese tone perception with congruent and incongruent auditory and visual materials for speakers of Mandarin Chinese and speakers of non-tonal languages. In general, we assume that (native and tone-naïve) participants mainly depend on auditory information when they have to identify Mandarin Chinese tones: both groups of participants are expected to identify the congruent stimuli more accurately than the incongruent ones, because (congruent) visual information can facilitate speech perception, especially for perceivers who lack comprehensive knowledge of the language (tone-naïve participants), while this additional value of visual cues would be less important for native participants.

When participants are presented with the incongruent experimental materials, we consider three types of possible outcomes: non-integration,

integration, and attenuation. For example, if the auditory input is mid-rising tone 2, but the visual input is high-level tone 1, and it turns out that the participant's percept is either tone 1 or tone 2, then this would indicate that perceivers choose to ignore the information in one channel and give preference to the other channel (non-integration). Another possible outcome is that the participants perceive a tone that is different from both tone 1 and tone 2 and that, consequently these cues were combined into a novel percept (e.g., a high-falling tone 4 or low-dipping tone 3); this would be a case of integration, as perceivers appear to combine the acoustic and visual channel and integrate them into a "new" tone. The third possible outcome would be a case where participants perceive a non-existing tone, whose height is between high (tone 1) and middle (tone 2) and the direction is in between rising and level (which we would call attenuation). Our current study will allow us to test whether the perceptual results can be explained in terms of integration or non-integration. It is not possible to differentiate between the first and third scenarios (attenuation), because of the nature of the experiment. With four obligatory response categories, participants still need to choose one of the two modalities, but their choice might be less certain. We expect that non-integration is most likely to happen for native Chinese participants (who are likely to ignore the visual channel), given that they can perfectly identify tones without seeing the speaker's face if the auditory information is clear. However, predicting the patterns that will emerge for the tone-naïve participants will be less straightforward. Given visual information would be more pronounced among tone-naïve participants, integration or non-integration are both likely to happen. The precise process might also depend on the difficulty tone-naïve participants have with identifying certain tones. In particular, it seems that the high-level tone 1 is more likely to be confused by inconsistent visual cues, as there are little or no visual activities in the nature of this tone. Since tone 3 is the mostly visually salient one (Mixdorff et al., 2005), it is expected that visual cues for tone 3 will exert the most influence on tone perception. Specific potential mixed patterns are expected to be found in the actual experimental results (which we present in the form of a confusion matrix).

4.2 Methodology

Two groups of participants (native Chinese and non-tonal language speakers) were tested with Chinese tone combinations of auditory-visual congruent stimuli (A_xV_x) and incongruent stimuli (A_xV_y). Accuracy, defined as the percentage correct identification of a tone based on its auditory realization, was used as the dependent variable.

4.2.1 Participants

A total of 142 participants comprised the two groups with different language backgrounds. The tone-naïve group consisted of 81 non-tonal language speakers (mean age 22, 49 female), mainly with a Dutch language background (n = 65). They were recruited from the participant pool for students of Communication and Information Sciences at Tilburg University. The other group consisted of 61 native speakers of Mandarin Chinese (mean age 25, 45 female) who were enrolled as students at Tilburg University, and they were recruited on campus. The participants either received 0.5 course credit for their participation or a gift card worth 5 euros.

4.2.2 Stimuli

A word list with 10 Mandarin monosyllables (e.g., *ma, ying* …) was constructed (based on stimulus material from Francis et al., 2008 and from Chen & Massaro, 2008, previously used by Chapter 2 and Chapter 3). Each of these syllables was chosen in such a way that the four tones would generate four different meanings, resulting in 40 (10 syllables × 4 tones) different existing words in Mandarin Chinese (see the Appendix 1 for a complete list of the stimuli).

A female native Mandarin Chinese speaker (age 31) produced the 40 words. She was given the instruction to "pronounce these words as if you were addressing someone who is not a Chinese speaker." There were no other instructions or constraints imposed on the way the stimuli should be produced. Every stimulus was pronounced twice. We used a Sony HDR-XR550VE camera to record the speaker's image and sound, resulting in one long video clip containing 80 words (40 words, each produced twice).

Windows Movie Maker (2018) was used to segment the long clip into individual tokens, with each token containing one syllable. All individual tokens last 2 seconds. We used Adobe Premiere Pro CC 2019 to create congruent and incongruent experimental stimuli by separating the image and the sound of one video into two channels and mixing the audio from one syllable with the image of the other. Care was taken to get precise synchronization between audio and video signals. These were aligned at syllable onset and the negligible perceptual temporal discrepancies at the syllable offset for incongruent tones were not discernable for our participants. In this way, for each stimulus, there are 12 incongruent combinations (A1V2, A1V3, A1V4, A2V1, A2V3, A2V4, A3V1, A3V2, A3V4, A4V1, A4V2, A4V3, where A refers to the audio channel and V to the video channel) and four congruent combinations (A1V1, A2V2, A3V3, A4V4). In order to ensure uniformity, the congruent stimuli were also cross spliced: for each stimulus, the image is taken from the first recorded

clip, and the sound from the second video clip. In total, 160 (10 syllables × 16 combinations) experimental stimuli were constructed.

In addition, to make sure that the participants would always attend the visual information, instead of focusing on the auditory channel alone, a 2-second silent video clip was created with a visible red dot on a still face (see Figure 4.1). When participants saw this red dot video, they had to press a designated button. Four of these video clips were mixed into those 160 tonal materials.[10]

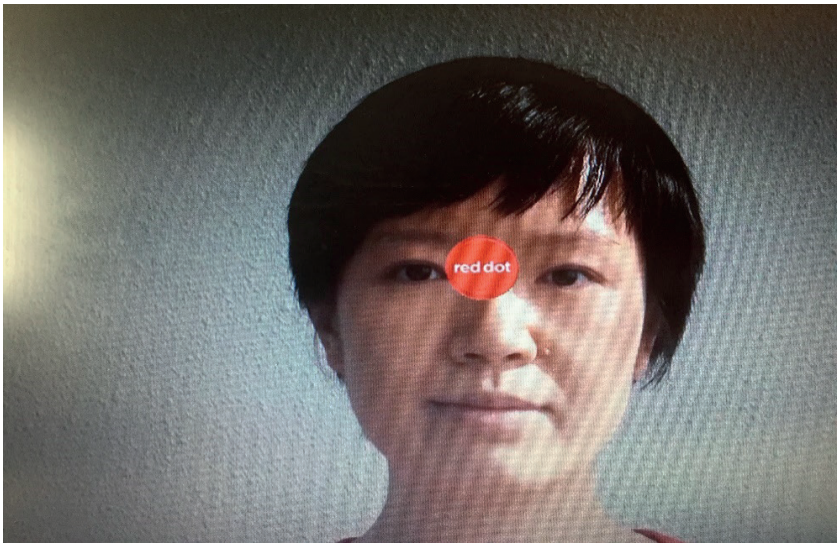


Figure 4.1. Screenshot of the red dot video

4.2.3 Procedure

All sessions were conducted in a sound-attenuated room. E-Prime 3.0 software (Psychology Software Tools, Pittsburgh, PA) was used to set up and run the experiment. The full procedure consisted of three blocks: instruction, practice trials, and test trials. Before the experiment started, participants were asked to fill out a questionnaire that assessed their language background in order to

[10]Data from participants who gave four wrong responses to the red-dot stimuli were excluded from the analyses. There were three of them in total.

be able to assign each participant to one of the participant groups (i.e., native speaker of Mandarin Chinese, native speaker of a non-tonal language). Native speakers of languages with tonal properties other than Mandarin Chinese were excluded from participation (there were four of them in total: two Norwegian, one Yoruba, and one Lithuanian). After that, a brief instruction about Mandarin Chinese tones was first displayed on the screen (see Figure 4.2): "there are four tones in Mandarin Chinese: the first tone is a High-Level tone, symbolized as " ˉ ", the second tone is a Mid-Rising tone, symbolized as "ˊ", the third tone is a Low-Dipping tone, symbolized as " ˇ ", and the fourth tone is a High-Falling tone, symbolized as " ˋ ".

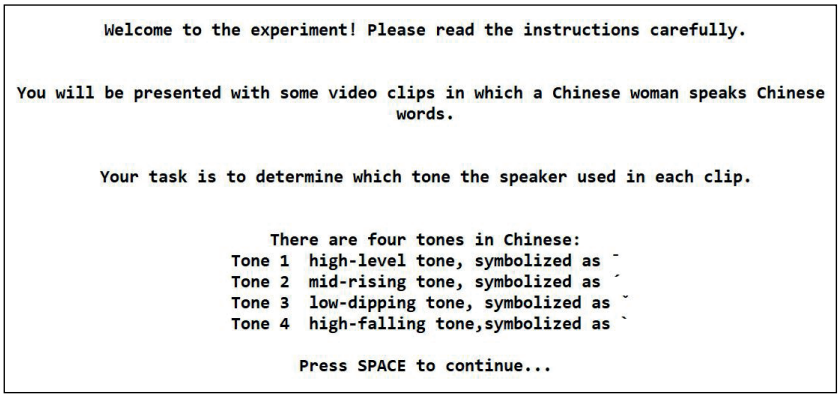


Figure 4.2. Screenshot of a brief introduction of Mandarin Chinese tones

The task of the participants was to identify the tones they perceived from the videos, written as "to determine which tone the speaker used". Six practice trials (five tonal video clips with a different speaker from the speaker in the test trial and one red dot clip) were included to familiarize participants with the testing procedure. After the practice trials, the experiment leader checked with the participants whether they had fully understood the concept of tones (in particular the symbols) and the task[11]. Then, the testing part of the study started (Figure 4.3 illustrates the testing path). The 164 test stimuli (160 tonal clips and four red dot clips) were presented in an individually randomized

[11]In a previous study (Chapter 2), we showed that tone-naïve participants have no problem to link pitch contour to visual and acoustic cues. In addition, many studies in the area of speech perception (e.g., Mixdorff et al., 2005; Burnham et al., 2001) have shown that perceivers will almost invariably use any reliable cue to facilitate their perception.

order (operated by E-Prime). The time for participants to give responses was 10 seconds, and there was no feedback (correct or wrong) given for their responses. Responses given outside the 10 seconds were treated as missing values.

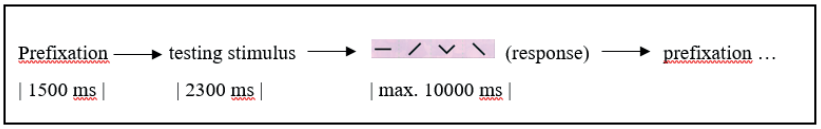


Figure 4.3. Time course of the testing stimuli

Participants wore headsets, and were seated directly in front of the PC running the experiment. All stimuli were presented at a comfortable hearing level. The participants were instructed to press the designated keys with the corresponding tone symbols and the red dot on them (“¯”, “ˊ”, “ˇ”, “ˋ” “RD”, see Figure 4.4) as accurately and as quickly as possible after they made their decisions. Their responses were recorded automatically by E-prime.

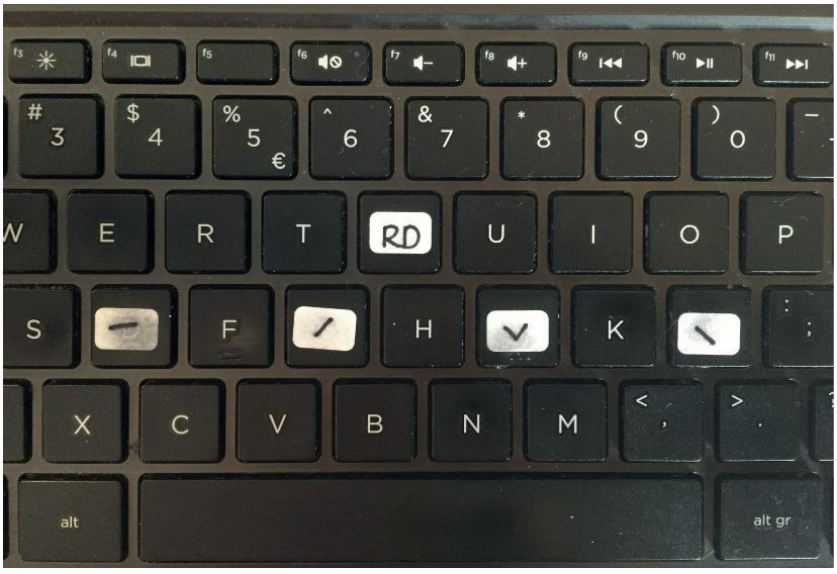


Figure 4.4. Picture of the designated keys with tone symbols and red dot (RD) on them

4.3 Results

This study is designed to investigate the perception of incongruent auditory-visual Mandarin Chinese tonal information. The experiment has a complete 2 × 2 design with congruency (congruent or incongruent) as the main within-subject factor and language background (native speakers of Mandarin Chinese or non-tonal languages) as the major between-subject factor. We included tone as another within-subject factor, and other factors, namely subject and syllable, were introduced as random factors. The results were analyzed by fitting linear mixed effects model (in R 3.6.0) for the dependent variable (the proportion of correct responses)[12] for both participant groups separately (Baayen, 2008) and by presenting confusion matrices for each auditory-visual combination. A correct response is defined as the proportion of correct identification of a tone based on auditory input. In addition, since there are four tones as the options, the basic chance of giving a correct response is 25%.

4.3.1 How would a McGurk effect work at the tone level for native speakers of Chinese?

To answer this first question, the performance of the 61 native Chinese participants was analyzed. Statistically, an effect of auditory-visual integration would be apparent in a main effect of congruency. To investigate this, it is necessary to incorporate random effects of subjects as well as syllables. In order to do so, we fitted a linear mixed effects model in R (version 3.6.0, R Core Team, 2019) using the package lme4 (Bates, Maechler, Bolker, & Walker, 2015). Following Barr, Levy, Scheepers, and Tily (2013), who recommend fitting a so called maximal model containing all random slopes and intercepts, we started out with a maximal model and removed random slopes until the model fit reached convergence. In our case, the first model that converged was a random intercept only model[13]:

1) Accuracy ~ Congruency * Tone + (1|Subject) + (1|Syllable), data = Chinese, family = "binomial"

This model fitted the data reasonably well (AIC = 568.8, log likelihood = -278.4), but did not yield a significant effect of congruency (β = 0.22, *SE* = 0.75, *z* = 0.30, *p* = .77), indicating that participants judged congruent stimuli (*M* =

[12]As many other previous McGurk effect papers (e.g., McGurk & MacDonald, 1976; Sekiyama, 1991, 1993, and 1997), we report accuracy as the dependent output in this paper, instead of both accuracy and speed, to address our research questions.

[13]As mentioned, more maximal models did not converge, but their parameter fit and significance was not meaningfully different from our chosen model.

0.994, *SD* = 0.071) equally well as incongruent stimuli (*M* = 0.995, *SD* = 0.067). The analysis did reveal a significant effect of tone (β = 0.48, *SE* = 0.16, *z* = 2.955, *p* = .003), reflecting small but statistically significant differences between (some of) the very high levels of performance for the individual tones (M_{tone1} = 0.995, SD_{tone1} = 0.070, M_{tone2} = 0.989, SD_{tone2} = 0.104, M_{tone3} = 0.997, SD_{tone3} = 0.057, M_{tone4} = 0.999, SD_{tone4} = 0.020). Finally, there was no significant interaction between congruency and tone (β = -0.05, *SE* = 0.33, *z* = -0.16, *p* = .87). Foreshadowing the discussions, we deem it quite likely that the absence of a significant effect is due to ceiling effects caused by the very high accuracy. This is less likely to happen in our sample of tone-naïve listeners.

Table 4.1 gives the correct responses for each tone as a function of the various AV combinations. The data in the confusion matrix shows that the Chinese participants did indeed perform very well in this tone identification task. Native speakers of Chinese had no difficulty identifying the tones in the discrepant stimuli. The scores in Table 4.1 indicate that the perception of native Chinese is totally driven by the auditory input. Accordingly, the additional visual information did not affect native speakers significantly, even when the visual cues do not match the auditory information.

Stimuli	Response Categories			
Auditory-visual component	Tone 1	Tone 2	Tone 3	Tone 4
A1V1	608	0	0	1
A1V2	603	6	0	0
A1V3	608	1	0	1
A1V4	609	1	0	0
A2V1	1	603	6	0
A2V2	0	603	7	0
A2V3	1	599	10	0
A2V4	1	608	1	0
A3V1	0	2	608	0
A3V2	0	2	608	0
A3V3	0	1	608	1
A3V4	0	2	608	0
A4V1	0	0	0	610
A4V2	0	0	0	610
A4V3	1	0	0	609
A4V4	0	0	0	610

Table 4.1. Stimulus combinations (*n* = 610 for each combination, minus incidental missing responses), definitions of response categories, and responses in each category for Chinese participants (correct responses are based on the auditory input)

4.3.2 How much do visual cues affect tone-naïve listeners in identifying Mandarin Chinese tones?

While the first set of analyses shows that visual cues did not significantly influence native Chinese in identifying Mandarin tones, we now focus on the performance of the 81 tone-naïve listeners (mainly Dutch) to see how they responded to congruent and incongruent stimuli. To answer this question, we again started to fit a maximal linear mixed effects model. The first model fit that reached convergence was[14] the following:

2) Accuracy ~ Congruency * Tone + (1 + Congruency | Subject) + (1 + Congruency |syllable), data = Dutch, family = "binomial"

This model showed significant effects of both independent variables and their interaction. The effect of congruency (β = 0.83, SE = 0.11, z = 7.55, p < .001) indicated that listeners judged congruent stimuli (M = 0.43, SD = 0.50) more accurately than incongruent stimuli (M = 0.36, SD = 0.48). When there is a discrepancy between visual cues and acoustic information (incongruent stimuli), listeners tend to rely more on the auditory input than the visual cues (M = 0.36 vs. M = 0.25; t (9719) = 15.05, p < .001). In addition, the significant effect of tone shows the difficulty our tone-naïve listeners have with tone 4 (M_{tone4} = 0.21, SD_{tone4} = 0.40) and how well they do with tone 3 (M_{tone3} = 0.55, SD_{tone3} = 0.50; M_{tone1}= 0.35, SD_{tone1} = 0.48; M_{tone2} = 0.41, SD_{tone2} = 0.49). Finally, these effects are qualified by a significant interaction between congruency and tone (β = -0.21, SE = 0.04, z = -5.28, p < .001), mostly indicating that the judgment accuracy of tone 3 and tone 4 is not affected much by congruency, but the accuracy of tone 1 and 2 judgments increases when the auditory and visual information are consistent.

The tonal confusion matrix (Table 4.2) might give more insight into the way they perceive Mandarin tones.

The data in Table 4.2 shows that the low-dipping tone 3 is the least confusing tone for tone-naïve participants (M = 0.56 and M = 0.55 for congruent and incongruent stimuli respectively, with M = .55 being the average of the three incongruent variations), while the high-falling tone 4 is the most commonly misidentified tone (M = 0.22 and M = 0.20 for congruent and incongruent stimuli respectively, with M = 0.20 being the average of the three incongruent variations). For the incongruent stimuli, tone 4 is mostly confused with the high-level tone 1 (M = 0.41 in congruent and M = 0.40 in incongruent stimuli, with M = 0.40 being the average of the three incongruent variations), although

[14]Compared to the previous section, there were more substantial differences between the converging models, most notably in the absent significance of the main effect of tone, indicating its dependence on interactions with syllable.

the confusions are not necessarily symmetrical: tone 1 was mostly confused with mid-rising tone 2 (M = 0.35 and M = 0.42 for congruent and incongruent stimuli respectively), rather than with tone 4. Tone 2 is mostly confused with tone 3 (i.e., when tone-naïve participants heard a rising tone 2, but saw a falling tone 4, they most likely perceived it as low-dipping tone 3), and tone 3 was most likely to be perceived as tone 4. Notably, for tone-naïve participants, not all the congruent stimuli were easier to identify than the incongruent ones (e.g., the accuracy for A_3V_3 was lower than for A_3V_2). As mentioned above, there is an interaction between tone and congruency: the congruency differently influenced the identification of individual tones: congruent visual information contributed more to the identifications of tone 1 and tone 2 than to tone 3 and tone 4.

Stimuli	Response Categories			
Auditory-visual component	Tone 1	Tone 2	Tone 3	Tone 4
A1V1	364	287	67	92
A1V2	242	388	108	69
A1V3	278	340	91	99
A1V4	266	304	101	139
A2V1	189	320	160	141
A2V2	95	412	189	113
A2V3	103	313	231	160
A2V4	119	284	239	168
A3V1	93	94	421	202
A3V2	49	125	464	172
A3V3	54	103	451	202
A3V4	51	90	452	217
A4V1	370	204	87	148
A4V2	295	269	102	144
A4V3	302	200	107	200
A4V4	329	208	98	174

Table 4.2. Stimulus combinations (n = 810 for each combination, minus incidental missing responses), definitions of response categories, and responses in each category for tone-naïve participants (correct responses are based on the auditory input)

4.3.3 What are the roles of congruent and incongruent visual information in tone perception?

Our results show that congruent stimuli are recognized more accurately than incongruent stimuli. However, this could be due to two effects (or a combination of them). An obvious first explanation is that perceivers benefit from additional congruent visual information, which increases the accuracy of their tone identification of congruent stimuli compared to that of incongruent stimuli. Alternatively, perceivers could be hampered by incongruent audio-visual information, making their identification less accurate compared to audio-visually congruent stimuli. In order to assess the contribution of visual information, we compared our current results with those of a previous study (Chapter 3) in which 43 different Dutch listeners judged the same stimuli (uttered by four speakers), but in an audio-only condition. If performance in the audio-only condition is worse than in the congruent audio-visual condition, that would be evidence for the first explanation, where congruent visual information aids tone identification. Alternatively, if performance in the audio-only condition is better than or similar to that in the congruent audio-visual condition, that would be evidence for the idea the incongruent visual information hampers performance.

As before, we fitted a linear mixed effects model with accuracy as the dependent variable and condition (audio-only versus (congruent) audio-visual) and tone as independent variables. Syllable and subject (and initially, speaker[15]) were introduced as random effects, and the first model that converged was as follows:

3) Accuracy ~ Condition * Tone + (1 | Subject) + (1 | Syllable), data = AV+AO, family = "binomial")

This model showed a (very) small effect of condition (β = 0.45, *SE* = 0.16, z = 2.77, p = .006), indicating that performance in the audio-visual condition was slightly better (M = 0.43, SD = 0.50) than in the audio-only condition (M = 0.42, SD = 0.49). This effect was quantified by a significant interaction between condition and tone (β = -0.17, *SE* = 0.04, z = -4.17, $p < .001$), showing that accuracy for tone 2 improves with additional visual information, while some tones are unaffected (tone 1 and tone 4), and tone 3 gets somewhat worse. Most of this is likely related to the inherent differences in classification accuracy of tones, as reflected by the main effect of tone (β = -0.12, *SE* = 0.02, z = -5.05, $p <$ .001). As before, tone 3 is the most accurately identified tone (M = 0.59, SD =

[15]Models with speaker as a random effect failed to converge due to the redundancy of speaker and condition. A model with speaker as a fixed effect was not significantly different from the model presented.

0.50) and tone 4 is the most difficult one to identify (M = 0.24, SD = 0.43), with the other two tones in between (M_{tone1} = 0.43, SD_{tone1} = 0.50, M_{tone2} = 0.44, SD_{tone2} = 0.50). While this result is strictly speaking compatible with the interpretation that visual information is helpful, it only marginally increases the accuracy, and this increase depends on the tone in question. On the other hand, these data provide counterevidence for the idea that adding visual information is harmful in itself.

4.4 Discussion and conclusion

In this study, we tried to answer two questions: firstly, whether a McGurk effect can also be discerned at the tone level in native speakers of Mandarin Chinese. Secondly, how visual information affects tone perception for native speakers and non-native (tone-naïve) speakers. To do this, we extended the existing body of auditory-visual integration (McGurk effect) studies to the suprasegmental level of Mandarin Chinese tones. When comparing the relative contribution of auditory and visual information during Mandarin Chinese tone perception in a noise-free condition with congruent and incongruent auditory and visual Chinese material for native speakers of Chinese and non-tonal languages (mainly Dutch), we found that visual information did not significantly contribute to the tone identification for native speakers of Mandarin Chinese and when there is a discrepancy between visual cues and acoustic information, (native and tone-naïve) participants tend to rely more on the auditory input than on the visual cues. Unlike the native speakers of Mandarin Chinese, tone-naïve participants were significantly influenced by the visual information during their auditory-visual integration, and they identify tones more accurately in congruent stimuli than in incongruent stimuli.

Strictly speaking, this study is different from the original McGurk study and the other studies that applied a McGurk effect to speakers of different languages (e.g., Sekiyama, 1997): instead of exploring consonant perception, we focused on tone identification and the visual cues that improve/alter the acoustic perception. This implies a shift from lip-reading (visual cues for consonants perception) to a focus on the whole face, head, and neck movements (visual cues for tone identification). However, this study is still one that investigates possible audio-visual interactions across tonal and non-tonal language speakers. The concept of the McGurk effect was applied to the way the experimental material was created: various tone combinations of the auditory and visual information were used.

The finding that native speakers of Mandarin Chinese mainly relied on the acoustic information of the input when the acoustic information is clear (no added noise or stimulus degradation) and that visual information neither improved nor hampered the tone identification for native Chinese speakers

(they identify the congruent stimuli equally well as the incongruent ones) implies that they were able to ignore the visual information, which is in line with our prediction of non-integration for native participants. However, the lack of integration we observed among native Chinese participants does not imply that there is no McGurk effect at the tone level. The absence of a significant visual effect could be due to ceiling effects caused by the very high accuracy. To avoid the emergence of such a ceiling effect, follow up experiments could use a degraded audio signal (as in Burnham et al., 2001) to show a potential fusion of auditory and visual channels in native speakers. Note that despite this experimental incentive to look at faces, Chinese subjects were forced to have a look at the face, due to the experimental set-up that included stimuli that required a visual task (identify red dots), and may have unlearned to pay attention to visual cues in the facial area. This may then still be consistent with the outcome of the Japanese speech processing experiment (Sekiyama, 1994), in which participants initially paid attention to the visual information and then separated it subsequently from the auditory information, because they sensed the discrepancy between the two channels. If that was indeed the case, then it would suggest that integration did in fact occur among native participants, but it was so early and fast that it could not be captured by our experiment (the participants have to wait to give their response after the stimulus is displayed). In connection with the issue of a ceiling effect in accuracy, it may be useful for future studies to also look at measures (e.g., reaction times) other than accuracy among the native speakers in order to detect a potential effect of visual cues (Chen, 2003; Ladd & Morton, 1997; Vanrell, Mascaró, Torres-Tamarit, & Prieto, 2013).

While native Chinese participants most likely ignored the visual information, tone-naïve participants identified more tones accurately when stimuli were congruent than with incongruent stimuli (in other words, they did take visual information into account in the tone identification task). However, we also found that tone-naïve participants, just as the native participants, relied more on auditory information than visual information when perceiving an unintelligible language (Mandarin Chinese), which is also in line with our hypothesis. The confusion matrix revealed some patterns for cases where tone-naïve participants were presented with incongruent experimental stimuli: whenever tone 1 was presented (i.e., the auditory input is tone 1) with incongruent visual cues (i.e., A1V2, A1V3, or A1V4), tone 2 was chosen as the answer most often in all three of the incongruent conditions. When tone 2 or tone 3 were in the auditory channel, tone-naïve participants gave their answers based on the auditory information (i.e., they picked tone 2 or tone 3 as their answer most often). When the auditory input was tone 4, the majority of the answers were tone 1. Therefore, for incongruent combinations in tone 1, we see one possible example of "non-integration" (as discussed in the introduction) giving preference to the visual channel (A1V2) and two examples

of "integration" (A1V3 and A1V4) where the new tone 2 occurred as the majority response. For tone 4, we see similar patterns with both "integration" (A4V1) and "non-integration" (A4V2 and A4V3) occurring. The responses for incongruent conditions in tone 2 and tone 3 paint a different picture, given that the most often picked answer was still tone 2 or tone 3. That is, we see examples of "non-integration" as participants seemed to rely on the auditory channel in these cases. These varied effects indicate that auditory-visual integration happened among tone-naïve participants, albeit not for all incongruent stimuli. The reasons why a certain tone mostly is confused with another specific tone could be various. For example, one of the possible reasons for choosing tone 1 whenever tone 4 was presented could be that tone 1 is perceived as a kind of default tone, with an unmarked configuration (e.g., a tone without a clear contour). Thus, when the participants experience difficulties grasping the changing pattern of the pitch, they tend to choose the default tone, since pitch height and pitch contour are not mastered in parallel (Wang et al., 2003). This can also be explained from a cross-linguistic perspective about the categorical nature of the perception of tone contrasts by speakers of tonal languages and speakers of nontonal languages (e.g., Hallé et al., 2004; but also see Krishnan, Gandour & Bidelman, 2010, for a different, more neurobiologically oriented perspective).

In addition, the tone confusion matrices revealed that the intrinsic characteristics of the tones appear to be the main contributors to tone identification. Tone 3 was the easiest tone for listeners to identify, irrespective of the visual information that had been added to the auditory information. Tone 4 was the most difficult one to correctly recognize. This is possibly due to their specific acoustic attributes — tone 3 has the longest duration and two intensity peaks, while tone 4 has the shortest duration, and only one intensity peak. Such features of the acoustic information have been preserved in the stimuli and they may have visual correlates as well (Mixdorff et al., 2005; Xu & Sun, 2002). For example, in the case of Mandarin tone 3 (low-dipping in terms of height and contour), the correlated head/neck motion during tone production should be signaled by a low-falling-rising movement. When present, these visual cues seem to be used by listeners during auditory-visual perception (Vatikiotis-Bateson et al., 2000), which has been shown by our finding of a significantly higher accuracy in the auditory-visual condition as compared to the audio-only condition. Such a result indicates that visual information helps tone-naïve participants to identify Mandarin tones. However, it only marginally increases the accuracy, and this increase depends on the tone in question: the accuracy of tone 3 and tone 4 is not affected much by congruency, but the accuracy of tone 1 and 2 judgments increases when the auditory and visual information are consistent.

Note also that the visual cues from the speaker in these videos are natural and, consequently, fairly subtle. We did not give extra instructions to the speaker about how to read out the Chinese words/tones, except for the instruction that she had to imagine addressing a foreigner. Native listeners have no difficulties recognizing the tones, which indicates that these recordings are unambiguous for them. On the other hand, our tone-naïve listeners do rely somewhat on visual information to assist their tone identification. In that case, salient visual information may better serve the purpose of testing congruent and incongruent visual information in their auditory-visual integration. Although the introduction in our experiment ("speak to a foreigner") to some extent already pushed the speaker to produce hyperarticulated speech, we realize that there is variation between speakers in terms of speech intelligibility: some speakers are easier to be understood than others (e.g., Cox et al., 1987; Ferguson, 2004) and the clarity of the visual cues they provide (e.g., Grant & Braida, 1991; Chapter 2). For future studies, it would be useful to employ multiple speakers to produce the stimuli, so that more hyperarticulated speaking styles could result in stronger incongruent visual information that could influence the native speakers, which would be favorable to a visual-only condition for native subjects.

In summary, native speakers of Mandarin Chinese who accurately identified the Chinese tones predominantly rely on auditory information of the input, even when incongruent visual information was present. Because of the high accuracy, a ceiling effect might have obscured auditory and visual integration among native Chinese participants, so the existence of a McGurk effect at the tone level cannot be entirely ruled out. Tone-naïve participants, on the other hand, were affected by visual information. However, while visual information is helpful for tone-naïve participants with incongruent stimuli, it only marginally increases the accuracy in the tone identification task compared to auditory information alone, and this increase depends on the tone in question. Relatively speaking, in a communicative context in which one can see the speaker's face, acoustic information contributed more for tone-naïve listeners in their tone identification as compared to visual information. In addition, identification varies with individual tones, with tone 3 (the low-dipping tone) the easiest one to identify, whereas tone 4 (the high-falling tone) was the most difficult one to perceive and tone 3 and tone 4 are not affected much by incongruency, but the accuracy of tone 1 and 2 judgments increases when the auditory and visual information are consistent.

Chapter 5

Chapter 5

Automatic Classification of Produced and Perceived Mandarin Tones on the Basis of Acoustic and Visual Properties

Abstract. This study addresses two questions. First, we investigate which acoustic and visual features of tones produced by native speakers can be used to automatically classify Mandarin tones. Second, we explore whether these features are similar to or different from the ones that have cue value for tone-naïve perceivers when they categorize tones. To address these questions we video-taped four Mandarin speakers while they produced ten syllables with four Mandarin tones, that is, 40 words in two styles (natural and teaching), totaling 160 stimuli. These audiovisual stimuli were subsequently presented to 43 tone-naïve participants in a tone identification task. Basic acoustic and visual features were extracted. We used Random Forest to identify the most important acoustic and visual features for classifying the tones, and Logistic Regression to train our classifiers on produced tone classification (given a set of auditory and visual features, predict the produced tone) and on perceived/responded tone classification (given a set of features, predict the corresponding tone as identified by the participant). The results showed that acoustic features outperformed visual features for tone classification, both for the classification of the intended and the perceived tone. However, tone-naïve perceivers did revert to the use of visual information in certain cases. So, visual information does not seem to play a significant role in native speakers' tone production, but tone-naïve perceivers do sometimes consider visual information in their tone identification.*

*This chapter is based on: Han, Y., Castro Ferreira T., Goudbeek, M., Mos, M., & Swerts, M. (Submitted). Automatic Classification of Produced and Perceived Mandarin Tones on the Basis of Acoustic and Visual Properties. Submitted for journal publication.

5.1 Introduction

Mandarin tones have been shown to have clear acoustic correlates, notably in the form of pitch and pitch contour. In particular, fundamental frequency (F0, perceived as pitch) patterns (both height and contour) and the direction of pitch, can distinguish four main distinctive tones, conventionally numbered 1 to 4: tone 1: high-level (5-5[16]); tone 2: mid-rising (or mid-high-rising; 3-5); tone 3: low-dipping (also low-falling-rising or mid-falling-rising; 2-1-4); and tone 4: high-falling (5-1) (Chao, 1930). Although tonal contrast in Mandarin Chinese is conveyed mainly by the height and contour of fundamental frequency (F0) (as the correlate of voice pitch (Francis et al., 2008; Hallé et al., 2004; Kong, 1987), Mandarin tones have also been shown to have other acoustic variables that can be perceptually informative (Chen & Massaro, 2008; Ryant et al., 2014), such as duration and amplitude (associated with intensity).

Tones vary systematically in duration in isolation: tone 1 and tone 4 tend to be shorter than tone 2, while tone 3 has the longest vowel length (Ho, 1976). Accordingly, duration differences between tones are perceptually distinctive. Mandarin speakers tend to enhance a tonal contrast that are relatively confusable (between tone 2 and tone 3) by using different lengths (time) (Blicher et al., 1990). Tones in Mandarin Chinese are differentiated by amplitude contour as well. It has been found that amplitude curves positively correlate with F0 (Ho, 1976, Whalen & Xu, 1992). Moreover, (Mandarin) listeners are able to identify tone 2, 3 and 4 fairly well solely based on amplitude contour (Whalen & Xu, 1992). In other words, amplitude contour is a reliable cue for tone identification. Most relevant studies have included these main acoustic features, that is, F0 and amplitude, for tone classification. However, our study will use additional sets of acoustic measures, the choice of which was inspired by findings of previous work, and supplemented with extra measures to explore their possible cue value.

Although tone perception mainly relies on auditory information (Burnham & Lau, 1998; Magnotti et al., 2015), there is consistent evidence that visual information plays a role in tone perception (e.g., Chen & Massaro, 2008; Chapter 2; Mixdorff et al., 2005a, 2005b; Shaw et al., 2014; Reid et al., 2015;). For example, based on visual information only, native speakers of Cantonese can distinguish Cantonese tones significantly above chance under certain conditions (Burnham et al., 2001). Similarly, Chen and Massaro (2008) found that the performance of native Mandarin speakers in visual lexical-tone identification is statistically better than chance. In addition, Han et al. (2018, 2019) showed that tone-naïve participants appear to benefit from visual information when identifying tones:

[16]Ibid., p. 4.

participants in an audio-visual condition distinguished Mandarin Chinese tones more accurately than those in an audio-only condition.

A number of studies have explored the nature and locus of the visual cues in tone production and perception, and revealed fairly reliable configurations of visual cues related to tone perception. For instance, Chen and Massaro (2008) found that visual tone identification improved significantly after the participants were taught to pay attention to the visual movements of the neck, head and mouth. Related to this, strong correlations between head/jaw movements and Fo were also observed by Vatikiotis-Bateson and Yehia (1996) and Yehia et al. (2002) (see also in Burnham et al., 2006; Attina, Gibert, Vatikiotis-Bateson, & Burnham, 2010). Eyebrow movements (Swerts & Krahmer, 2010; Kim, Cvejic, & Davis, 2014) and lip movements (Dohen & Loevenbruck, 2005; Dohen, Loevenbruck, & Hill, 2006; Attina et al., 2010) were reported to be associated with prosodic contrasts as well. More recent audiovisual research on tone languages from Burnham et al. (2019) added larynx motion (in addition to head motion) as a possible cue for Thai tone classification, and they found positive evidence that this type of motion is important for Thai tone production. Another recent study conducted by Garg, Hamarneh, Jongman, Sereno, and Wang (2019) found direct evidence that facial movements made during Mandarin tone production align with pitch trajectories. Specific visual cues in Mandarin tone production were clearly defined in their study and it was shown how those visual cues are associated with each of the four Mandarin tones: the downward and upward head and eyebrow movements respectively follow the dipping and rising tone trajectories; a lip closing movement is associated with the falling tone, and there are minimal movements for the level tone.

In general, it appears that the way tones are acoustically realized is often also visually signaled, because our mouths and faces need to gesture in a certain way to produce a given tone. Since more attention in past research has been paid to segmental information and less attention to the perception and classification of lexical-tone in research on audiovisual speech perception, evidence from the literature on the possible relationships between tone production and visible speech movements in tone-rich languages, such as Mandarin Chinese, is still sparse. In this study, we aim to find out whether visual information (i.e., facial expression and facial pose) contributes to the classification of Mandarin tones over and above the information provided in the acoustic signal. More specifically, our current study tries to shed light on which acoustic and visual cues can be used as automatic predictors of tones in (native) speakers' audiovisual productions. Furthermore, we want to know if these cues are also used by (tone-naïve) perceivers to classify tones.

We investigate both production and perception in our study as it is known from previous work that there is not necessarily a direct and tight relation between speech production and perception (e.g., Wang et al., 1999, 2003). For

instance, some acoustic variation, while a systematic and potentially good classifier, may not work well in perception, because it is below a perceptual threshold. In other words, by studying both production and perception we can look into what the perceivers pick up from what the speaker produces acoustically and visually. Moreover, because of our interest in the acquisition of tonal categories, we used tone-naïve participants for our tone identification experiment. By comparing automatic and human classification for Mandarin Chinese tones, the representativeness of our models as models of tone learning can be established. If both the performance and cues used by our algorithms are comparable to that of our participants, then our models could be viewed as representative of how people learn tones. In contrast, there may also be a considerable amount of difference between the two, as, in theory, there could be reliable predictors for tone production that are not included in the tone perception model. When such factors are perceptually accessible (i.e., listeners are actually able to hear/perceive them reliably) but not exploited for tone classification, then this could be a reason to explicitly teach foreign/second learners of a lexical tone language to pay attention to these factors.

So, this study aims (1) to find out which acoustic and visual features taken from native speakers' audiovisual recordings can be used by machine learning to classify Mandarin tones, and whether these cues are also used by tone-naïve perceivers to classify tones, and (2) to assess whether and how visual information (i.e., facial expression and facial pose) contributes to the classification of Mandarin tones over and above the information provided in the acoustic signal. More specifically, we applied this technique of tone classification both for produced tones and for perceived tones (obtained from tone-naïve participants). To answer these questions, we asked native Mandarin Chinese speakers to produce isolated words/tones and we later asked tone-naïve listeners to identify those tones. Acoustic and visual features were measured from the audiovisual signals, and their importance for tone classification was ranked by using a Random Forest classifier, which we applied both to predict (1) the speaker's intended tone production and (2) the tones as perceived by tone-naïve listeners. Furthermore, logistic regression was employed to predict the tones produced by the native Mandarin Chinese speakers and the responded tones by the tone-naïve perceivers based on acoustic-only, visual-only or combinations of acoustic and visual features, specifically to explore the relative contribution of visual information (over and above acoustic information) to tone classification for speakers and perceivers. Although tonal features have already been included in Chinese speech-recognition systems, most of them exclusively looked into relatively basic acoustic features and/or prosodic features (such as fundamental frequency (F0), duration and energy) for continuous Mandarin speech recognition (e.g., Chao, Yang & Liu, 2012; Chang, Zhou, Di, Huang, & Lee, 2000; Kalinli, 2011). This paper is one of those attempting to

include visual features into tonal modeling to classify Mandarin tones and to see the influence of visual characteristics from speakers on perceivers during the task of tone identification.

Given that tone production and perception in Mandarin Chinese greatly relies on auditory information (Burnham & Lau, 1998; Magnotti et al., 2015), we hypothesize that acoustic features would be ranked as more important than visual features in tone classifications of both produced tones (by native Mandarin Chinese speakers) and perceived tones (by tone-naïve perceivers). The order of the important acoustic features should be more or less the same in both classification tasks. More specifically, we expect to see fundamental frequency as the most important cue in classifying Mandarin tones, since Fo is the acoustical correlate of pitch. We also expect duration and intensity to be considered as important acoustic features, since they are reliable cues for tone identification, as mentioned in the above section. As for the other basic acoustic features, such as voice quality and formant frequencies, the importance of their roles or their contribution to tone classification is less clear, but potentially of importance, since voice quality can vary with pitch (e.g., Swerts & Veldhuis, 2001). As for the visual features, even the most promising facial features explored in the literature (for instance, head and neck movements, Chen & Massaro, 2008; eyebrow and lip movements, Grag et al., 2019) are not likely to appear at the top of the ranking table of tone classification for produced tones. However, some of the visual features could still outperform some of the acoustic features in the table of tone classification for perceived tones. After all, tone-naïve perceivers who lack sufficient tonal knowledge tend to consider visual features to facilitate their tone identification (Burnham et al., 2001; Chapter 2, 3, and 4). In terms of the features' importance ranking, we assume that acoustic features would be the best predictors for the classification of produced tones, and that combined model of acoustic and visual features should be (slightly) better than a model with acoustic features alone.

The paper is organized as follows. The method section consists of three parts. The first part contains a description of the tone production data as well as the extracted acoustic and visual features that are used for the first classification task. Second, we describe the tone perception study that provides the data for the second classification task. Third, a machine learning study is conducted based on these two data sets. The results section includes two parts. First, the importance ranking of the selected acoustic and visual features is shown for each of the classification models (production and perception). Second, the accuracies of tone classification models are compared with acoustic-only features, visual-only features and acoustic + visual features.

5.2 Corpus construction

Ten Mandarin monosyllables (e.g., *ma, ying* …) (based on stimulus material from Francis et al., 2008 and from Chen & Massaro, 2008, previously used by Chapter 2 and Chapter 3, see Appendix 1) were selected to compose a word list as the experiment material. The syllables were chosen in such a way that four tones would generate four different meanings, resulting in 40 (10 Syllables × 4 tones) different existing words in Mandarin Chinese. These 40 words were produced by four adult Mandarin-Chinese native speakers (two females and two males) who were born and raised in China and had come to the Netherlands for their graduate studies. The speakers were instructed to produce the stimuli in two different scenarios in sequence: a natural mode ("pronounce these words as if you were talking to a Chinese speaker") and a teaching mode ("pronounce these words as if you were talking to someone who is not a Chinese speaker"). The recording of the natural stimuli was done first, after which we recorded the teaching style stimuli. No other constraints were imposed on the way the speakers should produce the stimuli. There was a 20-minute break between the two recordings to counter fatigue.

We used Windows Movie Maker (2012) and Eye-catcher (version 3.5.1) to record the images and sounds of the speakers. Eye-catcher allows to easily capture the full-frontal images of the speakers' faces, because its camera is located behind the computer screen. This way, the setting is similar to a real face-to-face situation. Eventually, we generated two sets of 160 video stimuli (10 syllables × 4 tones × 4 speakers): one set for natural style, and one set for teaching style. We then divided these two long sets into individual tokens, with each token containing exactly one stimulus. We used version 3.9.5 of Format Factory to extract the sound from the videos[17].

Acoustic and visual features extracts

A textgrid command was used to extract the audio from the video recordings. We manually checked all the processed segments to make sure all the sounding excerpts were captured fully and correctly. After that, we automatically extracted the basic standard acoustic characteristics of the materials. For this study, Praat 6.0.33 (Boersma & Weenink, 2017) was used to measure the acoustic features for each of the 320 speech tokens (4 speakers x 40 words x 2 speaking styles). The default parameter settings for speech analysis in Praat were used (for instance, the standard pitch range of 75 - 500 hertz). The selection of the features was guided by the recommendation of The Geneva Minimalistic Acoustic

[17]The corpus here is constructed based on the stimuli produced in Chapter 2, and previously used in Chapter 3.

Parameter Set (GeMAPS) for voice research (Eyben et al., 2015), which is based on previous literature (also seen in a more recent study, Tupper, Leung, Wang, Jongman, & Sereno, 2020) and the theoretical significance of the features. Table 5.1 shows the extracted acoustic features and their definitions (based on the acoustic features used in Goudbeek & Scherer, 2010).

The durational parameter measures the total duration of the entire stimuli (excluding pre- and post-utterance silence). Since all stimuli were monosyllables and vowel types do not systematically influence tone identification (Chen & Massarro, 2008), we treated the total duration of the sounding segment as the tone length (time). *Pitch* is the perceived fundamental frequency (F0). The pitch features were directly obtained from Praat and all measures were computed over the voiced part. The minimum and maximum of the pitch were set at 5% and 95%, respectively. The mean and standard deviation were calculated based on the values between the minimum and maximum pitch. We also determined the 25th and 75th percentile of the F0 measure. The measures of *intensity* were derived in the same way as the pitch features and they were expressed in dB. The mean, standard deviation, minimum pitch, and maximum pitch were computed by querying the intensity contour in Praat. In addition, the intensity below 500 Hz and 1k Hz was extracted. Two voice quality features were also extracted**:** *jitter and shimmer*, which reflect the amount of irregularity of the signal in pitch and intensity, respectively. Since there were female and male speakers, *formants* were included as these measures are relevant for speaker identity, which we think might give more insights to the differences between speaking styles as well.

Acoustic features	Description
Duration	Total duration of the sounding segment in *s*
Pitch variations	
Pitch_mean	Mean pitch of the voiced part
Pitch_sd	Pitch standard deviation
Pitch_05(min)	Pitch contour 5th percentile
Pitch_95(max)	Pitch contour 95th percentile
Pitch_25	Pitch contour 25th percentile
Pitch_75	Pitch contour 75th percentile
Intensity variations	
Intensity_mean	Mean intensity in dB
Intensity_sd	Standard deviation of the intensity in dB
Intensity_min	Intensity (absolute) minimum in dB
Intensity_max	Intensity (absolute) maximum in dB
Intensity_05	Intensity contour 5th percentile in dB
Intensity_95	Intensity contour 95th percentile in dB
Intensity_500	The proportion of intensity below 500 dB
Intensity_1k	The proportion of intensity below 1k dB
Voice quality variations	
Jitter_local	This is the average absolute difference between consecutive periods, divided by the average period.
Jitter_absolute	This is the average absolute difference between consecutive periods, in seconds
Shimmer_dB	This is the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20.
Shimmer_local	This is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.
Formants variations	
F1	Center frequency and bandwidth of the first formant (in Hz).
F2	Center frequency and bandwidth of the second formant (in Hz).
F3	Center frequency and bandwidth of the third formant (in Hz).

Table 5.1. The acoustic features and their definitions

The Computer Expression Recognition Toolbox (CERT) was employed to automatically code facial expressions in the videos. CERT is a fully automatic, real-time software tool that estimates Action Units (AUs) from the Facial Action Unit Coding System (FACS) (Littlewort et al., 2011). Every facial expression can be described by the set of AUs that compose it. Action Units (AUs) were measured for each of the 320 video segments (i.e., each stimulus). As previous literature (Chen & Massaro, 2008) has asserted that the relevant visual cues for tone mainly depend on the (intensity of the) movements of the mouth, head/chin, and neck, 15 relevant action units (AU) and 3 head poses were selected. Table 5.2 shows a close-up of the 15 action units which have been included for the analysis. CERT also outputs estimates of the 3 head poses: yaw (the direction of shaking "no"), pitch (the direction of nodding "yes"), and roll (the in-plane rotation of the face).

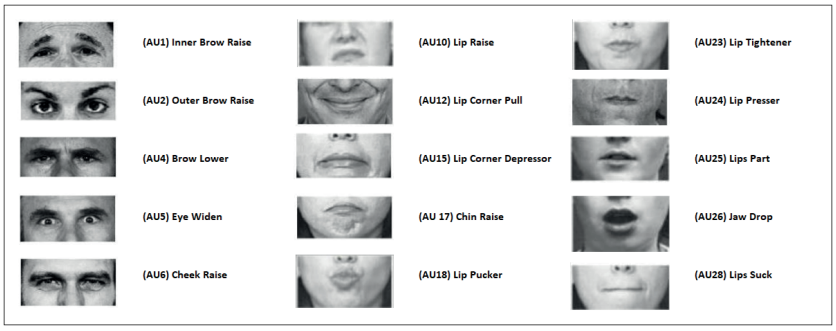


Table 5.2. Examples of the 15 selected Action Units from FACS (as cited in Cohn, Ambadar, & Ekman, 2007)

5.3 Perception study

43 participants (32 females and 11 males, with a mean age of 22) were recruited from Tilburg University. The majority of the participants (74%) were native speakers of Dutch. The 11 remaining participants each had a different native language: German, French, Greek, English, Portuguese, Spanish, Italian, Russian, Indonesian, Bengali, and Arabic. None of the participants were native speakers of tone languages, and no one was ever formally exposed to any tone language. The stimuli which were produced by the four native Mandarin Chinese were used to test the participants. All sessions were conducted in a

sound-attenuated room. E-prime (Version 2.0; Zuccolotto, Roush, Eschma, & Schneide, 2012) was used to set up and run the experiment[18].

The protocol of this study was approved by the Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences, Tilburg University. All participants gave written informed consent in accordance with the Declaration of Helsinki.

We found out that the accuracy of tone identification of the participants is well above chance (25%), $M = 48\%$, $SE = 0.02$. Tone 4 was the most difficult tone to recognize, with an accuracy of 32.1%. On the other hand, the low-dipping tone 3 was the easiest to recognize (62.7% accuracy). This can be seen in the confusion matrix (Table 5.3). The high-falling tone 4 was most often confused by the participants with the high-level tone 1. Tone 1 was mostly confused with mid-rising tone 2. Tone 1 and 2 showed little difference in accuracy, as the participants were able to identify them about equally (same as in Chapter 3 Table 3.1b).

		Responded tone			
		1	2	3	4
Presented tone	1	**49.8**	26.3	8.9	15.0
	2	16.8	**48.5**	20.7	14.1
	3	5.2	15.1	**62.7**	17.0
	4	34.7	23.0	10.3	**32.1**

Table 5.3. Confusion matrix for tone (percentage correct) in non-native perceivers

5.4 Machine Learning methods

To assess the importance of acoustic and visual features in the process of tone classification and tone perception, we made use of Machine Learning (ML), a subfield of Artificial Intelligence which aims to *learn* how to categorize data by using patterns and inference instead of explicit instructions. More specifically, we developed classification models of *Supervised Learning*, which are trained to predict an output *class* from input *features* based on examples of input-output pairs.

[18]The data in this perception study is the same data collected from non-musicians in Chapter 3.

5.4.1 Data

As we mentioned before, this study has two main goals. The first goal is to find out which acoustic and visual features (which are taken from native speakers' audiovisual recordings) can be used by machine learning to classify Mandarin tones. Furthermore, we investigate whether or not these cues are also used by tone-naïve perceivers to classify tones. The second goal is to see if visual information contributes to the classification of Mandarin tones over and above the information provided in the acoustic signal. If it does, we also want to see how it contributes to this classification. To be more specific, we applied this technique of tone classification for both the produced tones and for the perceived tones (all obtained from tone-naïve participants). Furthermore, to gain a better insight into the behavior of the participants in the tone perception experiment, we take a closer look at patterns in the errors they made. For instance, they might have made mistakes because they ignored the visual cues. To summarize, this leads to the following three classification experiments:

- Tone classification for produced tones

Given a set of features which describes the stimulus, the goal of the classifier was to predict the tones produced by the native Mandarin Chinese speakers. These classifiers were trained and evaluated based on 312 pairs of trial features and the corresponding tone extracted from production data.

- Tone classification for *all* perceived tones

Given a set of features which describes the stimulus and a participant, a tone perception classifier aimed to predict the corresponding tone guessed by the participant. 6702 pairs between trial and perceived tone were used to train and evaluate the classifiers.

- Tone classification for *wrongly* perceived tones

Given a set of features which describes the stimulus and a participant, a tone perception classifier was run to predict the wrong corresponding tone guessed by the participant. 3218 pairs between trial and wrongly perceived tone were used to train and evaluate the classifiers.

5.4.2 Features

In addition to the acoustic (see Table 5.1) and visual features (see Table 5.2) of each trial, the training models included a number of Common Features as

fixed factors: Syllable (10 levels), Speaker (4), and Speaking style (2), and for the perception data Perceiver (43).

5.4.3 Models and settings

Using the Scikit-learn framework (Pedregosa et al., 2011), we implemented two popular Machine Learning classifiers to answer our research questions: a *Random Forest classifier* and a *Logistic Regression*.

- The Random Forest classifier was used to rank the importance of each visual and acoustic feature in predicting the produced tone and the perceived one. In other words, this ranking classifier aims to find out which auditory and visual features are better cues for classifying Chinese tones and have more perceptual validity for tone-naïve learners of such tones, respectively. For each task, the classifier was trained with 100 estimators and default parameters in the framework.

- Logistic Regression was used to classify the produced and perceived tones. This technique has hyper-parameters, such as the normalization function and the parameter C, which needs to be tuned in order to fit the model to solve the classification problem. To do so, we used an optimization technique called Random Search (Bergstra & Bengio, 2012) to compare the l1- and l2-normalization functions and to consider 1, 5, 10, 50, 100, 500 and 1000 as possible values for C. Moreover, in order to speed up the training process, we pre-processed the feature inputs, scaling each of them to values between 0 and 1. Equation 1 presents this scaling process used for each feature:

$$std = (X - min(X)) / (max(X) - min(X))$$
$$X\_scaled = X\_std * (max - min) + min$$

Equation 1. Preprocessing method to scale the features between 0 and 1

5.4.4 Evaluation

With the proposed models, we performed two techniques: ranking and classification.

- Ranking

The proposed Random Forest approach was trained on each data set and was used to estimate the importance of each feature in the tone classification for produced tones and tone classification for perceived tones, respectively.

- Classification

Regarding the Logistic Regressions, we trained three classifiers for each task (e.g., tone classification for produced tones and tone classification for (all and wrongly) perceived tones) with the trials modelled by Common features + (1) acoustic features, (2) visual features, and (3) combined features (i.e., acoustic and visual features). In the target task, performance was evaluated through 10-fold cross validation, whereby the accuracies of the classifiers in the test sets of each fold iteration was averaged. The accuracy is the ratio of the number of correct predictions (i.e., true positives and true negatives) over the total testing samples.

By training and measuring the performance of these classifiers, we aim to find out the importance of each type of feature in the task of tone classification. Since we assume that the acoustic-only model will outperform the visual-only model and visual features may help distinguishing the produced and perceived tones, we expect that the classifiers which model the trials with combined features will have a higher performance than the acoustic- and visual-only classifiers.

5.5 Results

We ranked all the selected acoustic and visual features predicting produced tones and perceived tones and reported the accuracy of each tone classification model in this section. Linear regression was conducted to see how well the selected acoustic and visual features in predicting tones. The overall model fit for produced tones was $R^2 = .61$. The model fit for all perceived tones was $R^2 = .13$ and for wrongly-perceived tones was $R^2 = .20$.

5.5.1 Ranking

Table 5.4 and 5.5 depict the importance of each feature in the tasks of tone classification for produced tones and for *all* perceived tones, respectively. The importance value of all the features should sum up to 1, and the average importance is the cutting point (sum/number of features). In this case, the cutting point of the importance is 0.02. Features which have greater value than the cutting point should be considered as important and should be included in the model.

As shown in Table 5.4, for classifying the tones produced by native speakers, acoustic features in general were ranked more important than visual features. Duration was the most important feature to categorize the tones (0.09), followed by jitter_local and pitch_05 (0.06). Pitch variations including pitch_75 (i.e., pitch contour 75th percentile), pitch_25 (i.e., pitch contour 25th

percentile), and pitch_mean, and voice quality parameter, jitter_absolute, took the third position (0.05). Among all the visual cues, only motions from lips (lips_suck and lips_corner pull; 0.02) and eyebrows (0.02) were relevant cues for classifying produced Mandarin Chinese tones, the others were not considered as important. Three potential head poses (pitch, yaw, and roll) were not seen as important visual features in predicting produced tones as expected. Some of the features representing variations of intensity (mean and minimum of the intensity, for instance) and voice formants were ranked as being of low importance (0.01). Among all the common features, only speaker should be considered as important, while syllable and speaking style only had an importance of 0.01.

Rank	Feature Name	Importance	Feature Type
1	Duration	0.09	acoustic
2	Jitter_local	0.06	acoustic
	Pitch_05	0.06	acoustic
3	Pitch_75	0.05	acoustic
	Jitter_absolute	0.05	acoustic
	Pitch_mean	0.05	acoustic
	Pitch_25	0.05	acoustic
4	Pitch_sd	0.04	acoustic
	Shimmer_local	0.04	acoustic
	Pitch_95	0.04	acoustic
	Shimmer_dB	0.04	acoustic
5	Intensity_max	0.02	acoustic
	Speaker	0.02	common
	Intensity_05	0.02	acoustic
	Intensity_95	0.02	acoustic
	Intensity_sd	0.02	acoustic
	Intensity_500	0.02	acoustic
	AU1_Inner_Brow_Raise	0.02	visual
	AU12_Lip_Corner_Pull	0.02	visual
	AU28_Lips_Suck	0.02	visual
6	F1	0.01	acoustic
	AU25_Lips_Part	0.01	visual
	AU4_Brow_Lower	0.01	visual

	AU2_Outer_Brow_Raise	0.01	visual
	Intensity_mean	0.01	acoustic
	Intensity_min	0.01	acoustic
	AU6_Cheek_Raise	0.01	visual
	AU17_Chin_Raise	0.01	visual
	Pitch	0.01	visual
	F3	0.01	acoustic
	Yaw	0.01	visual
	AU26_Jaw_Drop	0.01	visual
	F2	0.01	acoustic
	Roll	0.01	visual
	AU10_Lip_Raise	0.01	visual
	AU18_Lip_Pucker	0.01	visual
	AU15_Lip_Corner_Depressor	0.01	visual
	AU24_Lip_Presser	0.01	visual
	Intensity_1k	0.01	acoustic
	AU5_Eye_Widen	0.01	visual
	AU23_Lip_Tightener	0.01	visual
	Syllable	0.01	common
	Speaking style	0.01	common
Sum	Common features	0.04	
	Acoustic features	0.76	
	Visual features	0.20	

Table 5.4. Tonal features ranking in tone classification for produced tones. Note that visual cues are indicated in grey.

Rank	Feature Name	Importance	Feature Type
1	Perceiver	0.75	common
2	Duration	0.03	acoustic
3	Pitch_05	0.01	acoustic
	Shimmer_local	0.01	acoustic
	Jitter_absolute	0.01	acoustic
	Pitch_25	0.01	acoustic
	Jitter_local	0.01	acoustic
	Pitch_sd	0.01	acoustic
	Pitch_75	0.01	acoustic
	Pitch_mean	0.01	acoustic
	Shimmer_dB	0.01	acoustic
	Pitch_95	0.01	acoustic
	AU1_Inner_Brow_Raise	0.01	visual
	Speaker	0.01	common
	Intensity_95	0.01	acoustic
	Intensity_sd	0.01	acoustic
	AU12_Lip_Corner_Pull	0	visual
	AU2_Outer_Brow_Raise	0	visual
	AU28_Lips_Suck	0	visual
	AU4_Brow_Lower	0	visual
	AU25_Lips_Part	0	visual
	F2	0	acoustic
	AU18_Lip_Pucker	0	visual
	F1	0	acoustic
	Intensity_05	0	acoustic
	AU6_Cheek_Raise	0	visual
	Intensity_max	0	acoustic
	AU10_Lip_Raise	0	visual
	Pitch	0	visual
	AU26_Jaw_Drop	0	visual
	Yaw	0	visual
	F3	0	acoustic
	AU15_Lip_Corner_Depressor	0	visual
	Intensity_500	0	acoustic

	Intensity_min	0	acoustic
	AU5_Eye_Widen	0	visual
	Syllable	0	common
	AU17_Chin_Raise	0	visual
	Intensity_1k	0	acoustic
	Roll	0	visual
	AU23_Lip_Tightener	0	visual
	AU24_Lip_Presser	0	visual
	Intensity_mean	0	acoustic
	Speaking style	0	common
Sum	Common features	0.76	
	Acoustic features	0.17	
	Visual features	0.01	

Table 5.5. Tonal features ranking in tone classification for *all* perceived tones. Note that visual cues are indicated in grey.

Table 5.5 displays the results of the feature ranking in the task of tone classification for *all* perceived tones. The amount of important tonal features for classifying perceived tone was considerably less than the number of features involved in predicting the produced tones (16 vs. 43). There were only two features in total considered as important features: perceiver and duration. Perceiver, with an important value of 0.75, was the most important feature to classify the perceived tones, which indicates some participants were better than the others in the task of identifying Mandarin tones. Similarly to the production classification, duration again turns out to be a very important feature, with an importance of 0.03. Pitch variations, jitter, shimmer, and eyebrow motion appeared to play small but equal roles in tone classification for all perceived tones. Generally, however, it appears that no particular visual cue contributed significantly to the classification.

Rank	Feature Name	Importance	Feature Type
1	Perceiver	0.68	common
2	Duration	0.02	acoustic
3	Pitch_05	0.01	acoustic
	Shimmer_local	0.01	acoustic
	Jitter_absolute	0.01	acoustic
	Pitch_25	0.01	acoustic
	Jitter_local	0.01	acoustic
	Pitch_sd	0.01	acoustic
	Pitch_mean	0.01	acoustic
	Pitch_75	0.01	acoustic
	Shimmer_dB	0.01	acoustic
	Pitch_95	0.01	acoustic
	Speaker	0.01	common
	AU2_Outer_Brow_Raise	0.01	visual
	AU18_Lip_Pucker	0.01	visual
	Intensity_05	0.01	acoustic
	Intensity_sd	0.01	acoustic
	AU1_Inner_Brow_Raise	0.01	visual
	AU28_Lip_Suck	0.01	visual
	Intensity_max	0.01	acoustic
	AU4_Brow_Lower	0.01	visual
	AU12_Lip_Corner_Pull	0.01	visual
	F3	0.01	acoustic
	F2	0.01	acoustic
	Intensity_min	0.01	acoustic
	AU25_Lips_Part	0.01	visual
	F1	0.01	acoustic
	Intensity_95	0.01	acoustic
	AU24_Lip_Presser	0.01	visual
	Roll	0.01	visual
	AU15_Lip_Corner_Depressor	0.01	visual
	Intensity_500	0.01	acoustic
	Intensity_1k	0.01	acoustic
	AU23_Lip_Tightener	0.01	visual

	Yaw	0.01	visual
	AU6_Cheek_Raise	0.01	visual
	Pitch	0.01	visual
	Intensity_mean	0.01	acoustic
	AU5_Eye_Widen	0.01	visual
	Syllable	0.01	common
	AU26_Jaw_Drop	0.01	visual
	AU10_Lip_Raise	0.01	visual
	AU17_Chin_Raise	0	visual
	Speaking style	0	common
Sum	Common features	0.70	
	Acoustic features	0.23	
	Visual features	0.17	

Table 5.6. Tonal features ranking in tone classification for *wrongly* perceived tones. Note that visual cues are indicated in grey.

Table 5.6 provides a ranking of the tonal features for classifying the incorrect responses from the perceivers, i.e., those responses where the tone that was identified by the participant was not the one uttered in the stimulus. Perceiver was again ranked as the most important feature (0.68), just as in the *All* Perceived Tones ranking (Table 5.5), followed by duration (0.02). Compared to the feature ranking of tone classification for all perceived tones, there were more features involved when perceivers responded incorrectly (16 vs. 39). However, similar to the previous findings, even when perceivers gave incorrect responses, they still paid more attention to the acoustic cues than to the visual cues.

5.5.2 Tone classification

Table 5.7 shows the averaged general and tone-specific accuracies of the Logistic Regression classifiers in the tasks of tone classification for produced tones and tone classification for (all and wrongly) perceived tones. On average, we see that the acoustic-only model outperformed the visual-only model in the classifications for produced tones and perceived tones. Concerning the *combined* version of our models, the results reveal that combining acoustic and visual features did not lead to a better classification than the *Acoustic-only* model. These results are consistent with the features' importance ranking,

showing that acoustic features are the best predictors for the classification of Mandarin Chinese tones, and that the modeled visual features are not significant predictors for the classification tasks.

Tone classification for produced tones					
	General	Tone 1	Tone 2	Tone 3	Tone 4
AO	0.82 (0.78-0.86)	0.84 (0.73-0.94)	0.68 (0.56-0.80)	0.90 (0.84-0.95)	0.88 (0.78-0.99)
VO	0.20 (0.16-0.24)	0.22 (0.11-0.33)	0.18 (0.10-0.27)	0.09 (-0.02-0.21)	0.32 (0.18-0.46)
AV	0.83 (0.79-0.87)	0.83 (0.74-0.93)	0.73 (0.62-0.83)	0.89 (0.76-1.01)	0.91 (0.86-0.96)
Tone classification for *all* perceived tones					
	General	Tone 1	Tone 2	Tone 3	Tone 4
AO	0.50 (0.48-0.52)	0.62 (0.60-0.64)	0.45 (0.42-0.48)	0.63 (0.60-0.66)	0.22 (0.19-0.25)
VO	0.32 (0.31-0.32)	0.40 (0.37-0.44)	0.48 (0.45-0.51)	0.23 (0.20-0.26)	0.05 (0.03-0.07)
AV	0.50 (0.48-0.52)	0.63 (0.61-0.65)	0.44 (0.41-0.46)	0.64 (0.62-0.66)	0.22 (0.19-0.25)
Tone classification for *wrongly* perceived tones					
	General	Tone 1	Tone 2	Tone 3	Tone 4
AO	0.44 (0.42-0.46)	0.58 (0.56-0.61)	0.51 (0.48-0.54)	0.23 (0.17-0.28)	0.33 (0.28-0.38)
VO	0.36 (0.34-0.37)	0.35 (0.33-0.38)	0.63 (0.60-0.66)	0.08 (0.06-0.11)	0.20 (0.16-0.23)
AV	0.48 (0.46-0.50)	0.62 (0.59-0.64)	0.51 (0.48-0.53)	0.30 (0.27-0.34)	0.42 (0.36-0.47)

Table 5.7. General and tone-specific accuracies in tone classifications. Note that numbers in brackets represent 95% confidence intervals.

Moreover, the data in Table 5.7 show that of the three tone classification tasks, the produced tones are the best fit according to our models. The reason for this is that the accuracies of the acoustic-only model and the combined model in tone classification for produced tones were much higher than the corresponding accuracies in both all and wrongly perceived tones (82% in AO and 83% in AV for produced tone classification vs. 50% in both AO and AV for perceived tone classification). Table 5.7 also tells us that there is no difference in performance between AO and AV models overall. However, the model of visual-only fits better in predicting perceived tones than in produced tones: the accuracy of the VO model is higher in tone classification for all perceived tones (especially for wrongly perceived tones) than in tone classification for produced tones (32%, 36% and 20%, respectively).

Although the visual-only models did not significantly contribute to the accuracies of the combined models in general, the influence of visual cues had a

different picture on tone classification on individual tone level. Zooming in on specific characteristics of individual tone classification for produced tones, we observe that the accuracy of a visual-only model for tone 4 (high-falling tone) (32%) was above chance (25%), indicating that in the visual-only condition, visual features can significantly contribute to classify tone 4. Similarly, in tone classification for all perceived tones, the accuracies of a visual-only model for high-level tone 1 and mid-rising tone 2 were also above chance (40% and 48%, respectively). Interestingly, the visual-only model was the best model to predict tone 2 in the tone identification task (48% in visual-only vs. 44% in combined). In addition, in the perceived tone classification, the acoustic-only model and the model based on combinations of acoustic and visual features made accurate predictions for high-level tone 1 (63%) and low-dipping tone 3 (64%), while all models (i.e., acoustic-only, visual-only, and combined) made poor predictions for high-falling tone 4 (22%, 5% and 22% respectively).

In the task of tone classification of wrongly perceived tones, the combined acoustic and visual model (AV) generated the highest degree of (predictive) accuracy (48% in AV; 44% in AO; and 36% in VO). In contrast to the other two classification tasks, for classifying *wrongly* perceived tones, including both auditory and visual features provides the best model. In other words, visual features seem to significantly contribute to the participants' (erroneous) performance, although acoustic features were still the most relevant predictors. This is the case for tone 2 in particular: the visual-only model (with an accuracy of 63%) is the most accurate model to predict tone 2, instead of the combined model (with an accuracy of 51%).

5.6 Discussion

In this study, we set out to answer two questions: first, which acoustic and visual cues presented in native producers' audiovisual signal can be used to automatically classify Mandarin tones and whether these cues are also used by tone-naïve perceivers to identify tones. Second, whether and how visual information (i.e., facial expression and facial pose) contributes to the classification of Mandarin tones over and above the information provided in the acoustic signal. In doing so, we video-taped four native Mandarin Chinese speakers producing 40 existing words in teaching and natural mode and then presented these videos to 43 tone-naïve participants to identify the tones. Basic acoustic and visual features were extracted from the experimental materials produced by the native speakers. We used random forest classifier to estimate the importance of the selected features and rank them in order of importance. Three logistic Regression classifiers were trained for each of the tone classification tasks (i.e., tone classification for produced tones by native speakers, tone classification for all and wrong perceived tones), to find out the

importance of each type of feature in the task of tone classification. In addition, we investigated whether the features used to classify the produced tones are similar to or different from the ones that tone-naïve perceivers use when they identify tones.

First, the results of the linear regression showed that the selected acoustic and visual features can be used efficiently in the automatic tone classification models. Second, the ranking tables of the tonal features showed that most of the acoustic features were ranked as more important than the visual features in Mandarin Chinese tone classification for the intended/produced and the perceived tones (which is in line with our assumption). Among all the chosen acoustic parameters (i.e., pitch, intensity, voice quality, duration and formants), duration stood at the top of the ranking in produced tone classification and was ranked as the secondary important feature (after perceiver) in perceived tone classification. In other words, duration is the most important and reliable feature to classify Mandarin Chinese tones. Pitch variations and voice quality (jitter and shimmer) were ranked below duration. Third, none of the selected visual features, not even the three head poses which were considered as the most promising visual cues, were ranked as important in tone classification for perceived tones. However, three visual features (AU1: inner brow raise, AU12 lip corner pull, and AU28 lips suck) were considered as important for classifying produced tones by native Chinese speakers.

The fact that pitch was not ranked as the most important feature in tone classification for perceived tones is not completely consistent with our hypothesis. Pitch, as the perceived form of fundamental frequency, was expected to be the most important feature for tone classification. A possible explanation for this might relate to the inadequate tone knowledge of the perceivers. Compared to duration (the length of the tone), pitch is a relative abstract feature since it involves tone direction and height. Because of its complexity and the time pressure in the experiment, our tone-naïve participants may have focused primarily on the length difference of the tone, which is more direct and easier to grasp.

Our analysis showed that the factor "Perceiver" was found to be the factor responsible for the largest amount of variance explained in the responses by our tone-naïve participants, indicating the importance of individual differences in tone perception. Because individual variation was responsible for so much variance, the contribution of other features might have been suppressed in the rankings. While we did address individual variation in previous work (Chapter 3), it was not the focus of this study, and future work could attempt to clarify the role of individual characteristics in the perception of Mandarin tones.

We found that the amount of useful tonal features for (all) *perceived* tone classification was much less than for *produced* tone classification. This result may be explained by the fact that the model is better at predicting what tone

was produced (r^2 = .609), than at what tone was perceived (r^2 = .133). It may well be that the features that the model considers map better to what is produced (i.e., these are in fact the relevant features to distinguish tones), than to which features are considered for perception (i.e., the possibility of the participants paying attention to other things than the variables that were included in the model), and therefore possibly relating to a mismatch between production and perception. Another possible explanation could be that the participants were in fact often genuinely guessing, and were randomly pressing one of the buttons. This latter explanation could also be the reason why individual variation was ranked as the most important factor in classification for perceived tones. The fact that some participants are better than others might be due to individual differences in perceptual abilities. For instance, one of our previous studies found that tone-naïve perceivers who have official musical experience identified Mandarin Chinese tones better than their counterparts who do not have musical experience (Chapter 3).

In line with the feature importance ranking, the results of models for tone classification showed that acoustic features are the best predictors for the classification of most Chinese tones, and visual features did not play a significant role in tone classification in general. This finding partially supports our assumption that an acoustic-only model would be the best predictor for tone classification for produced tones, while a combined model would be the best model of predicting the perceived tones by tone-naïve perceivers. Notably, the accuracy of the visual-only model in the (all and wrongly) perceived tone classification was higher than the one in produced tone classification, both for the general result and for each individual tone. This indicates that when native Chinese speakers produce tones, they barely use visual cues to convey the tonal information, while tone-naïve perceivers do consider the visual cues when they give (wrong) responses. In other words, in the task of tone identification, visual information misled tone-naïve participants to some extent. However, even when visual cues did not necessarily lead the participants to a higher accuracy of identification, the effect may appear from other dependent variables (not addressed here), such as shorter reaction times. Therefore, monitoring reaction time could be the pursuit of future work.

Based on the data of predicting wrongly perceived tones, we also noticed that visual features had some influence on tone classification for a subset of cases, particularly for mid-rising tone 2 in the tone perception experiment. The accuracy of identifying tone 2 was improved by 3% (from 45% to 48%) when moving from an acoustic-only model to a visual-only model. Interestingly, the accuracy of the model even became worse when the acoustic-only model was combined with visual-only. This suggests that it is more likely to get tone 2 correctly identified by the perceivers when they only focus on the visual cues and ignore the acoustic information. Although the role of visual features for tone

2 was not so dominant in tone classification for the produced tone, the accuracy of the model combining auditory and visual features was higher than that with only auditory features (from 68% to 73%). The amount of visual information involved in the production of individual Mandarin Chinese tones varies between tones (Chen & Massaro, 2008). Interestingly, the low-dipping tone 3 has the most visually observable activity, which could lead to the prediction that tone 3 is the tone for which visual information matters most. However, the visual-only model not showing a high accuracy on predicting tone 3 may be caused by the fact that it has unique and strong acoustic properties (the longest duration and two intensity peaks), which may have helped perceivers to easily identify it, even without considering visual cues. All models (i.e., acoustic, visual and combined) were bad in predicting tone 4, which implies that tone 4 happens to be the most difficult tone for identification among all the Mandarin tones, so neither the acoustic nor the visual features are really helpful for perceivers, in line with our results from previous studies (Chapter 2 and Chapter 4). This could be due to the unique and specific acoustic attributes of tone 4 (i.e., the shortest duration, and only one intensity peak), while very brief visual activities make tone 4 difficult to perceive.

5.7 Conclusion

Facial features have been mostly neglected in tonal classification. Most studies to date have focused on the importance of the role of acoustic features for automated speech recognition. This makes sense, since visual features do not significantly contribute to the improvement of tone classification models, as our study showed. However, visual features should still be included in the model of tone classification when the model is to predict perceivers' behavior. For Mandarin Chinese tone classification, acoustic features weigh more than visual features, and duration should be considered as the most important acoustic feature for native speaker's speech recognition. Visual features are especially helpful in identifying the mid-rising tone 2. In addition, for tone-naïve speakers, high-falling tone 4 is the most difficult tone to recognize.

Chapter 6

Chapter 6

General Discussion and Conclusion

The previous chapters presented four studies on factors influencing tone perception, specifically the effect of visual information on Mandarin Chinese tone identification by tone-naïve perceivers, in combination with other contextual and individual factors. The relative importance of acoustic and visual information for tone perception and (automatic) tone classification have been explored as well. In this final chapter, I will first summarize the main findings and formulate answers to the main research questions that were introduced in Chapter 1. Next, theoretical implications of these findings will be discussed, as well as directions for future research.

6.1 Main findings

The first research question, investigated in chapters 2 and 3, was whether modality (audio-visual vs. audio-only condition), combined with speaking style (natural vs. teaching mode) and musicality of the perceivers (musicians vs. non-musicians), affects the perception of Mandarin Chinese tones by tone-naïve speakers. The tone identification experiment presented in Chapter 2 showed that the video conditions (audio-visual natural and audio-visual teaching) resulted in an overall higher accuracy in tone perception than the auditory-only conditions (audio-only natural and audio-only teaching), but no better performance was observed in the audio-visual conditions in terms of reaction time, compared to the auditory-only conditions. The finding that participants in the audio-visual conditions outperformed their counterparts in the audio-only conditions supported our claim that tone-naïve perceivers can benefit from visual cues that native speakers display on their faces. Teaching style turned out to make no difference on the speed or accuracy of Mandarin tone perception (as compared to a natural speaking style). Therefore, the hypothesis that speaking style would influence Mandarin tone perception was rejected. In chapter 3, we presented the same experimental materials and procedure, but now with musicians versus non-musicians as participants. The results provided further evidence for the view that the availability of visual cues along with auditory information is useful for people who have no knowledge

of Mandarin Chinese tones when they need to learn to identify these tones. The data also revealed that musicians outperformed the non-musicians in both experimental conditions (audio-visual and audio-only), which was in line with our hypothesis that musical ability positively affects the ability to identify Mandarin tones.

Apart from the factors we focused on (i.e., modality, speaking style and musicality), we explored two additional factors in chapters 2 and 3 that could potentially affect the ability to perceive Mandarin Chinese tones, even when these factors were not our primary interest. First, there is variation between speakers. We investigated whether some speakers speak in such a way that tones are easier to identify than those from other speakers. Second, some tones may be easier to perceive than others. We particularly assessed whether the way the tones are acoustically realized is also visually signaled, which could lead to differences in accuracy of tone recognition (Chapter 2). The analyses revealed differential effects of speaker on the accuracy of the tone identification. That is, some (female) speakers produce tones that are easier to identify than the tones of other (male) speakers. Regardless of conditions, the specific tone had a striking influence on the percentage of correct responses and on reaction time as well. Specifically, we found that tone 3 (low-dipping tone) was the easiest one for the listeners to identify, while tone 4 (high-falling tone) was the most difficult one.

In addition to the findings mentioned above, Chapter 3 also revealed that musicians and non-musicians equally improved their tone identification accuracy over time, which showed both groups were learning at the same rate. Although musicians performed better than non-musicians across the board (they had a higher accuracy overall), they did not learn faster than non-musicians. Musicians' overall accuracy was higher to start with, but did not increase more than the accuracy of non-musicians. Moreover, musicality also predicted accuracy for each of the individual tones. Out of the five sub-scales (active engagement, perceptual abilities, musical training, singing abilities, and emotions) measured by Goldsmiths Musical Sophistication Index, musical training was the only constant factor predicting the accuracy for each individual tone, while the other factors had no consistent effect on accuracy. This suggests that implementing musical training could facilitate (tone) language learning, and that is promising for educational purposes.

So far, the results of the first two experiments presented in chapters 2 and 3 showed that adding visual cues to clear auditory information facilitated the tone identification for tone-naïve perceivers (there is a significantly higher accuracy in audio-visual condition(s) than in auditory-only condition(s)). This visual facilitation did not change with the presence of (hyperarticulated) speaking style or the musical skill of the participants. Moreover, variations in

speakers and tones had effect on the accurate identification of Mandarin tones by tone-naïve perceivers.

The next two studies, reported on in subsequent chapters, focused on the relative importance of auditory and visual information in tone perception for tone-naïve perceivers (Chapter 4) and automatic tone classification (Chapter 5). More specifically, in Chapter 4, we tried to answer whether or not there is an audio-visual integration at the tone level in native speakers of Mandarin Chinese and tone-naïve perceivers (i.e., we explored perceptual fusion between auditory and visual information, reminiscent of the well-known McGurk effect). We found that visual information did not significantly contribute to the identification of Mandarin Chinese tones. When there is a discrepancy between visual cues and acoustic information (such as when the auditory input is mid-rising tone 2, but the visual input is high-level tone 1), both native and tone-naïve participants tend to rely more on the auditory than on the visual information. In other words, our findings provide no evidence of auditory-visual fusion for tone perception among native speakers of Mandarin Chinese. Unlike native speakers, tone-naïve participants were *influenced* by the visual information: they identified the tones more accurately in congruent stimuli than in incongruent stimuli. In addition, in line with our previous findings, individual characteristics of the tones appear to be the main contributors to accurate tone identification. Tone 3 was the easiest tone for listeners to identify, irrespective of the visual information that had been added to the auditory information. Tone 4 was the most difficult one to correctly recognize.

Finally, Chapter 5 investigated which acoustic and visual properties from the native speakers' audio-visual signal can be used to automatically classify Mandarin tones and whether these features are similar to or different from the ones that are used by tone-naïve perceivers when they categorize tones. The results showed that acoustic features outperformed visual features for tone classification, both for the classification of the intended/produced and the perceived tone. However, tone-naïve perceivers did revert to the use of visual information in certain cases (i.e., when they misjudged tones). Thus, visual information does not seem to play a role in native speakers' tone production, but tone-naïve perceivers do sometimes consider visual information in their tone identification. These findings provided additional evidence that auditory information is more important than visual information in Mandarin tone perception and tone classification. Notably, visual features contributed to the participants' erroneous performance. This suggests that visual information actually misled tone-naïve perceivers in their task of tone identification. To some extent, this is consistent with our claim that visual cues do influence tone perception. In addition, the ranking of the auditory features and visual features in tone perception showed that the factor perceiver (i.e., the participant) was responsible for the largest amount of variance explained in the responses by

our tone-naïve participants, indicating the importance of individual differences between perceivers in tone perception.

To sum up, perceivers who do not have tone knowledge in their language background tend to make use of visual cues from the speakers' faces for their perception of unknown tones (Mandarin Chinese in this dissertation), in addition to the auditory information they clearly also use. Thus, auditory cues are still the primary source they rely on. There is a consistent finding across the studies that the low-dipping tone 3 is the easiest tone to identify, while the high-falling tone 4 is the most difficult one to recognize.

6.2 Theoretical implications

This section discusses implications of the findings for existing theories of audio-visual tone perception. More specifically, I will indicate where the present research supports existing theories and where it diverges from them, and I will also briefly touch upon the necessity of studying individual differences between perceivers in tone perception.

6.2.1 Audio-visual tone perception

One of the central points in this dissertation is the value of visual information in tone perception by tone-naïve perceivers. The findings of the studies presented here paint a mixed picture with respect to this issue. In Chapter 2, our finding that tone-naïve participants were (slightly) better able to identify tones when they *saw* the speaker than when they only *heard* them suggested that visual information plays a facilitating role in tone perception. Similar results were found in Chapter 3: musicians and non-musicians (tone-naïve perceivers) were able to identify tones better in audio-visual conditions than audio-only conditions. Although the effect was not that large, visual information *was* helpful for perceivers to identify Mandarin tones. In Chapter 4, the results showed that visual information marginally increased the accuracy in the tone identification task, but that this increase depended on the tone in question. Therefore, these first chapters showed positive effects of visual information on tone perception for tone-naïve perceivers. However, the modelling approach from Chapter 5 that attempted to predict the perceived Mandarin tones by tone-naïve perceivers revealed no significant differences between auditory-only and auditory-visual models. Instead, the results indicated that visual features seem to significantly contribute to the (tone-naïve) participants' erroneous performance in Mandarin tone perception. In other words, in the task of tone identification, visual information misled tone-naïve participants, at least to some extent.

Many previous studies were not able to reveal an audio-visual facilitation effect in *native* speakers except in noisy or impaired listening conditions (e.g., Burnham et al., 2015; Mixdorff et al., 2005a; Mixdorff & Charnvivit, 2004; Smith & Burnham, 2011). One explanation could be that the use of native speakers as participants may have obscured possible beneficial effects of the visual modality. A noisy or difficult listening situation pushes the native speakers to make use of the visual information, which may be underused by normal-hearing tone perceivers (Reid et al., 2015). In our pretest of Chapter 2 (see section 2.2.2), we also observed that native participants in the audio-only condition already performed at ceiling. In contrast to these previous findings, we did find a difference between audio-only and audio-visual conditions with "clear" stimuli (i.e., no noisy environment, degraded signal or hearing-impaired participants), but only for non-native listeners in Mandarin Chinese tone perception. In that respect, our study is in line with studies that have shown that visual information is available to, and used by, language learners during speech perception, and tone perception is determined by both auditory and visual information (Reid et al., 2015; Smith & Burnham, 2012; Burnham et al., 2015).

With respect to the question of relative strength of auditory and visual information in Mandarin tone perception, the results are consistent throughout our thesis: compared to visual information, auditory information contributes to a much larger extent to the accurate identification of tones, both for tone-native and tone-naïve perceivers. The tone identification experiment conducted with tone-native and tone-naïve perceivers with auditory-visual congruent and incongruent materials in Chapter 4 showed that native Chinese participants most likely ignored the visual information (they identify the congruent stimuli equally well as the incongruent ones), while tone-naïve participants did take visual information into account in the tone identification task (they identified more tones accurately when stimuli were congruent than with incongruent stimuli), but this only marginally increased their perceptual accuracy. That provided explicit evidence that auditory information played a dominant role in tone perception. Automatic classification for Mandarin Chinese tones presented in Chapter 5 was in line with this outcome, but also identified the relative importance of individual acoustic features. Our findings are thus in line with previous studies that indicate that tone perception (and tone production) in Mandarin Chinese greatly relies on auditory information (Burnham & Lau, 1998; Magnotti et al., 2015; Sekiyama & Burnham, 2008). Moreover, when comparing native speakers of (Mandarin) tone languages with native speakers of non-tonal languages in the task of tone perception, our findings provide support for the view that visual speech information appeared to be ignored by tone language speakers (Mandarin Chinese speakers, at least), while non-tonal language speakers seemed to be prone to perceptual fusion like effects at the

tone level, as they tend to make use of available visual cues and integrate them with the auditory information.

6.2.2 Individual differences between perceivers

In Chapter 5, "perceiver" was found to be the factor responsible for the largest amount of variance explained in the responses by our tone-naïve participants, indicating the importance of individual differences in tone perception. The foundation of individual difference research is that it examines attributes on which learners vary and how such variations relate to language-learning success. The individual variables among learners that can have an impact on the learning process and have been studied plenty in previous literature include: learning aptitude, gender, culture, age, learning styles, learning strategies, and affective variables (see, for instance, Dörnyei & Skehan, 2003; Ehrman, Leaver, & Oxford, 2003; Schmidt, 2012; Tagarelli, Ruiz, Vega, & Rebuschat, 2016). Given the design of our experiments (preliminary data and short time span), in this dissertation, I focus on looking into the individual differences between perceivers in tone perception, instead of relating the individual differences to more general aspects of second language learning.

In Chapter 3, we addressed one perceiver related characteristic that clearly influences tone identification, as it was shown that musicality (in particular, musical training) positively contributes to Mandarin tone identification. There is a large literature concerning the correlation between musical ability and tone learning (Delogu et al., 2006, 2010; Marie et al., 2011; Ong et al., 2017; Wong & Perrachione, 2007) and there is evidence from previous studies that musical training facilitates the learning of linguistic tone (e.g., Alexander et al., 2005). Our finding that musicians are at an advantage compared to non-musicians when identifying tone in Mandarin Chinese is in line with previous studies. Additionally, we looked into the performance of musicians and non-musicians over time in order to see whether the two groups differ in learning rate. The learning patterns showed that although musicians showed their superior performance at the beginning of the task, they did not learn faster than non-musicians. Interestingly, the increase in performance follows a linear pattern for both musicians and non-musicians, and does not seem to plateau, indicating that more exposure leads to better performance, and potentially (in the case of a longer learning period) may lead to still higher final accuracy scores. It therefore appears that for people with no tone language experience, musical training aids in linguistic tone perception. However, a deeper insight into the relationship between musical training and tone identification in musicians is still lacking. If there would be replications in the future or similar work focuses on this relationship, we would suggest incorporating a behavioral or cognitive test (e.g., Woodcock-Johnson Tests of Cognitive Abilities WJ-Cog; Woodcock,

1997) to better characterize the perceptual differences and relate them to tone identification.

The other individual characteristic from perceivers that had an obvious effect on tone identification is the language background of the two groups of participants: tone-native and tone-naïve perceivers. In Chapter 4, we presented various tone combinations of congruent and incongruent auditory-visual materials to native speakers of Mandarin Chinese and speakers of tone-naïve languages. The results showed that both groups relied mainly on auditory information (rather than visual information) when perceiving Mandarin Chinese. If we would have pursued this further, we would have tried to unravel the details of the perceptual mechanism that distinguishes tone perception in tone-native and tone-naïve speakers. However, we could draw some inspiration from previous studies about the variabilities in ability to successfully learn to use pitch in lexical contexts. For instance, Gandour (1983) suggested that different language backgrounds may be associated with variability in cue-weighting, i.e., speakers of a contour tone language, such as Mandarin Chinese, tend to attend more to a dimension related to pitch *direction*, while non-tone speakers tend to place more emphasis on pitch *height* (speaker-specific information) (see also Li & Shuai, 2011). Since the major predictor of successfully learning to use lexical tones is the ability to perceive pitch contours (Wong & Perrachione, 2007; Moreno et al., 2009), this correlates well with the fact that some participants outperformed others, even though all of them were non-tone speakers, because good perceivers/learners attended more to pitch direction compared to poor perceivers/learners (Chandarsekaran, Sampath, & Wong, 2010).

Taken together, it is fair to say that tone perception can vary considerably between individuals, and that taking individual differences in tone learning into consideration will maximally benefit all perceivers.

6.2.3 A theory of tone perception

There is a consistent finding across the studies in this dissertation that certain tones are perceived more easily than others, which could relate to potentially universal aspect of tone perception mentioned by Burnham et al. (2015), also in as far as visual cues are concerned. In their experiment, they found that (for Thai) the tone *contour* provides the best information across modalities and that articulation of dynamic tones perhaps has more obvious visual concomitants than static tones. A similar result was found in their study in 2001 on Cantonese tone perception: dynamic tones have more obvious visual concomitants than static tones (in visual-only mode). Moreover, among these contour tones, rising tones are more salient than falling tones (Burnham et al., 2001; 2015). As for Mandarin Chinese tones, there is one static tone (Tone 1) and three contour tones (Tone 2, Tone 3 and Tone 4). Specifically, across the studies in

this dissertation, we have found that low-dipping tone 3 was the easiest tone to identify for tone-naïve perceivers, while high-falling tone 4 was the most difficult one. The tone confusion matrices (Chapter 2, 3, and 4) showed that tone 4 was most commonly misidentified as high-level tone 1, while tone 1 was mostly confused with mid-rising tone 2. In addition, tone 2 appeared to be the one influenced the most by visual cues from the speakers.

Although the pattern is quite consistent across our studies, the conclusion is not entirely in line with the findings from previous studies: while tone 3 has indeed been found to be the easiest one to identify (due to the longest vowel duration), tone 4 has not been clearly found to be the most difficult one to recognize (Mixdorff et al., 2005b; Blicher et al., 1990; Fu & Zeng, 2000). Moreover, the confusion patterns reported in various studies seem to be influenced by the linguistic background of the perceivers. So (2006) examined Mandarin tone identification by native speakers of Cantonese and Japanese, and found that Cantonese speakers confused more often between Mandarin Tone 1 and Tone 4 (the same as we found in tone-naïve speakers), while Japanese speakers confused Tone 2 and Tone 4 more often. Ultimately, a common confusing pair for all three groups (Cantonese, Japanese and English) is Mandarin Tone 2 and Tone 3 (So & Best, 2011).

Mandarin tones are intrinsically different from each other, acoustically and visually. For instance, tone 3 has the longest duration and two intensity peaks, while tone 4 has the shortest duration, and only one intensity peak. With respect to visual features, tone identification has been found to mainly depend on the (intensity of the) movements of the mouth, head/chin, and neck: specifically, there is little to no activity for tone 1, some activity for tone 2 and tone 4 (though very brief for tone 4), with tone 3 having the most activity, namely a dipping head/chin. Duration (time) differences between the tones may be caused by variation in the movements of the mouth, as more complex movements would require more time to be realized (Chen & Massaro, 2008). Taken together, it is much more important *which* tone the listeners hear than *how* they hear it (e.g., whether visual information is present).

6.3 Suggestions for future work

First of all, we suggest that future work could benefit from pursuing the investigation of individual factors, both of speakers and perceivers, in the process of tone perception. Which individual characteristics eventually lead to efficient communication is one of the questions that this dissertation has not fully answered. Since the results in Chapter 2 indicated that female speakers were easier to be understood by the tone-naïve perceivers than male speakers, the gender of the speakers was identified as a possibly mitigating factor. However, more data, especially with a sufficiently large sample of speakers, should be

collected to broadly reveal the speakers' variations in tone identification, as there were only four speakers in our study, which does not allow us to attach much importance to it. For example, an acoustic analysis for each speaker's speech (tone) production would be useful, so that specific auditory properties could be identified as the responsible factors for the easiness of perception. Similarly, for visual information, further research should take into account the variation between speakers' realizations of visual information. After all, there might be substantial differences in the degree to which the speakers' faces exhibit relevant characteristics. In addition to individual differences between speakers, individual differences between perceivers should also be considered in future work on tone perception. In Chapter 5, perceiver was found to be the factor responsible for the largest amount of variance explained in the responses by our tone-naïve participants, indicating the importance of individual differences in tone perception. Future work can pursue incorporating a behavioral or cognitive test to be able to better characterize the differences in perceptual abilities and relate them to tone identification.

Second, most of the studies of this dissertation used accuracy as the dependent variable. However, we also included reaction time in the study presented in Chapter 2 to see the effect of modality and speaking style on tone identification for tone-naïve perceivers. Our analysis did not reveal many significant results in terms of reaction time (except for the finding that perceivers may react faster to a specific speaker). We therefore did not include reaction time in any later studies, since using accuracy was sufficient to answer our research questions. However, as we stated in the discussion sections of chapter 4 and 5, we would like to encourage future work to look at other measures as well, including reaction times. The reason for this is that sometimes accuracy cannot paint the whole picture: Reaction times can provide information about the difficulty or ease of the perception process, as reflected in the speed with which it happens. Even if the outcome (correct or incorrect tone identification) is the same, visual information might, for instance, facilitate the process in terms of processing speed, which can be captured in reaction times. Therefore, the results of reaction time and accuracy experiments do not always yield the same conclusions.

For instance, including the measure of reaction times could add instant value to our study presented in Chapter 4. We presented congruent and incongruent auditory-visual tone stimuli to native speakers of Mandarin Chinese and non-tone language speakers, and we found native Mandarin Chinese perceivers identified the congruent stimuli equally well as the incongruent ones. However, by focusing on accuracy only, we could not discern whether there is an auditory-visual fusion at the tone level due to ceiling effects caused by the high accuracy. Adding the measure of reaction times could remove the ceiling effect problem, as according to Prinzmetal, McCool & Park (2005), auditory or visual

channel selection does not affect the perceptual representation (performance), but involves a decision as to which location should be responded to. If there is a conflict as to which location should be responded to, responses are delayed (like they are in the incongruent stimuli situation). In other words, while visual cues did not necessarily lead the participants to a higher accuracy of tone identification, they might have led to shorter reaction times.

Third, there are some methodological implications related to the simultaneous study of production and perception. As mentioned in the introductory chapter of this dissertation, the assumption behind doing this was that we then can look into what the perceivers pick up from what the speaker acoustically and visually produces. In Chapter 5, we used computational models to directly investigate the relationship between the production and perception of tones, by modelling the categorization of the produced tones by native speakers and the perceived tones by tone-naïve perceivers. The comparison of the results showed that visual information does not seem to play a significant role in native speakers' tone production, but tone-naïve perceivers do sometimes consider visual information in their tone identification. This indicates that there is some discrepancy in the process of perceiving tones during the transition from tone production to tone perception. We believe that studying a combination of tone production and perception would provide added value to current and future studies. An immediate suggestion for future work would be looking into the production of tones by (tone-naïve) learners. In our studies, tone-naïve participants learned to identify tones, but not to produce them. An analysis of produced tones would provide new or additional insights into the relationship between tone production and tone perception.

Finally, the way the experimental stimuli were created could be improved by using a more natural setting. In this dissertation, the experimental stimuli (tones) were produced by native speakers of Mandarin Chinese in a natural and teaching style, but in imaginary scenarios (e.g., teaching style is the style "as if you were talking to someone who is not a Chinese speaker"). So while our elicitation was partly naturalistic, there were also clear elements of posed speech. We have further tried to tackle the issue of ecological validity in a number of ways. For example, we use Eye-catcher (version 3.5.1) to record the speakers' images and sounds. One of the advantages of Eye-catcher is that the camera is located behind the computer screen, which is convenient for capturing the full-frontal images of speakers' faces unobtrusively, similar to what listeners would see in a face-to-face setting (Chapter 2. 3 and 5). Nevertheless, future work could benefit from an actual natural or teaching context or from extracting stimuli from a relevant corpus (e.g., Modern Spoken Chinese Corpus (MSCC)). However, while we strongly encourage the use of more natural dialogue settings for stimulus elicitation, there is also evidence that even when no addressee is present, speakers still show a form of audience-designed linguistic behavior

(e.g., Koolen, Gatt, Goudbeek, & Krahmer, 2011; Van der Wege, 2009). For instance, in the paper of Uther, Knoll and Burnham (2007) that addressed the issue of 'foreigner-directed-speech' (FDS), one of the manipulations in the study was to contrast natural speech and a 'teaching' register, produced "as if you were talking to someone who is not a Chinese speaker". The results suggested that linguistic modifications found in foreigner-directed speech are didactically oriented. Therefore, it remains an open question how much of the effects attributed to the presence of an addressee are due to the physical presence of an addressee.

In addition, given the set-up of our stimulus recording (with elicited productions of isolated words), it could have been the case that our natural condition already represents rather "clear" speech, and in that sense is not representative of the reduced speech samples one often observes in more spontaneous data (Berry, 2009). Because of this, the tone may have been comparatively easy for listeners to distinguish. Accordingly, visual cues from the speaker in these videos are natural, but consequently fairly subtle. Follow-up experiments could use a degraded audio signal (as in Burnham et al., 2001) to show a potential fusion of auditory and visual channels in native speakers.

Furthermore, it is fair to raise some concerns about the generalizability of our findings in this dissertation. First of all, we only used isolated words as the experimental stimuli. Although this is a reasonable starting point for tone perception as it becomes complicated on a sentence level due to the involvement of assimilation or intonation, generalization on sentence and context level is a matter for speculation. However, tone perceptual studies on sentence context level have been done (e.g., Hallé et al., 2004; So & Best, 2011). For example, Burnham, Ciocca and Stokes (2001) found that (Cantonese) participants generally performed better with words in isolation than with words in context in the task of (Cantonese) tone identification, but isolation or context interacted with the presenting modes (AO, AV and VO). Second, while most of the present research has been conducted using university students, tone identification should also be studied in other groups of participants to see whether the results would generalize to language use outside the lab. Such work can be found as well (Cienkowski & Carney, 2002; Tye-Murray, Sommers, & Spehar, 2007; Wang et al., 2020). For instance, the most recent study was conducted on 6-and 8-year-old monolingual or bilingual children to make a systematic comparison of Mandarin tone training in the AO and AV modes. These findings provided evidence that tone language experience (monolingual or bilingual) is a strong predictor of learning unfamiliar tones (Kasisopa, El-Khoury Antonios, Jongman, Sereno, & Burnham, 2018). Hopefully, further work would benefit from the issues brought forward in the present dissertation.

6.4 General conclusion

The primary aim of this dissertation was to gain insight into the contribution of visual information to tone identification for tone-naïve perceivers. Much traditional work on tone perception has been a purely auditory phenomenon, which may be influenced by contextual, linguistic or conceptual factors. The studies presented here have shed light on the influence of visual information on tone perception. At the same time, they have also made it clear that there are differences in the effects of visual information on tone-native and tone-naïve speakers, in the relative cue value of auditory and visual speech information, and that the individual tones differ in learnability. These findings are important pieces of evidence that should be incorporated into both theoretical and application-oriented models of tone learning. More generally, we hope that our findings will benefit theories of audio-visual speech perception with data for tone and will inspire future research on second language studies of Mandarin Chinese tones.

References

Aheadi, A., Dixon, P., & Glover, S. (2010). A limiting feature of the Mozart effect: listening enhances mental rotation abilities in non-musicians but not musicians. *Psychology of Music*. *38*(1):107-117. https://doi.org/10.1177/0305735609336057

Alexander, J. A., Wong, P. C., & Bradlow, A. R. (2005, September). Lexical tone perception in musicians and non-musicians. In *Ninth European Conference on Speech Communication and Technology*.

Attina, V., Gibert, G., Vatikiotis-Bateson, E., & Burnham, D. (2010). Production of Mandarin lexical tones: Auditory and visual components. In *Auditory-Visual Speech Processing 2010*. Retrieved from: https://www.isca-speech.org/archive/avsp10/av10_S4-2.html

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

Baayen, R.H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. New York, NY: Cambridge University Press.

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23-36. https://doi.org/10.1016/j.jml.2015.10.008

Bailly, G., Perrier, P., & Vatikiotis-Bateson, E. (Eds.) (2012). *Audiovisual speech processing*. New York, NY: Cambridge University Press.

Bao, Z. (1990). *On the nature of tone* (Doctoral dissertation). Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/14143/23903579-MIT.pdf?sequence=2.

Bao, Z. (1999). *The structure of tone*. Oxford University Press, USA.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bates, D., Maechler. M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior and Development, 36*(4), 847-862. https://doi.org/10.1016/j.infbeh.2013.09.001

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(1), 281-305.

Berry, J. (2009). Tone space reduction in Mandarin Chinese. *The Journal of the Acoustical Society of America*, *125*(4), 2571-2571. https://doi.org/10.1121/1.4783753

Besson, M., Chobert, J., & Marie, C. (2011). Transfer of training between music and speech: common processing, attention, and memory. *Frontiers in psychology*, 2, 94. https://doi.org/10.3389/fpsyg.2011.00094

Besson, M., Schön, D., Moreno, S., Santos, A., & Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restorative neurology and neuroscience*, *25(*3-4), 399-410.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of speech and hearing research*, *17*(4), 619-630. https://doi.org/10.1044/jshr.1704.619

Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of Syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics*, *18*(1), 37-49. https://doi.org/10.1016/S0095-4470(19)30357-2

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.33, retrieved from http://www.praat.org/.

Bovo, R., Ciorba, A., Prosser, S., & Martini, A. (2009). The McGurk phenomenon in Italian listeners. *Acta Otorhinolaryngologica Italica*, *29*(4), 203.

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America, 121*(4), 2339-2349. https://doi.org/10.1121/1.2642103

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America, 112*(1), 272-284. https://doi.org/10.1121/1.1487837

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English/r/and/l: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299-2310. https://doi.org/10.1121/1.418276

Burnham, D., Brooker, R., & Reid, A. (2015). The effects of absolute pitch ability and musical training on lexical tone perception. *Psychology of Music*, *43*(6), 881-897. https://doi.org/10.1177/0305735614546359

Burnham, D., Ciocca, V., Lauw, C., Lau, S., & Stokes, S. (2000). Perception of visual information for Cantonese tones. *In Proceedings of the Eighth Australian International Conference on Speech Science and Technology*. pp. 86-91. Canberra: Australian Speech Science and Technology Association.

Burnham, D., Ciocca, V., & Stokes, S. (2001). Auditory-visual perception of lexical tone. I *Seventh European Conference on Speech Communication and Technology*. Retrieved from https://www.isca-speech.org/archive/eurospeech_2001/e01_0395.html

Burnham, D., & Dodd, B. (1996). Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. In *Speechreading by humans and machines* (pp. 103-114). Springer, Berlin, Heidelberg.

Burnham, D., & Dodd, B. (2018). Language–General Auditory–Visual Speech Perception: Thai–English and Japanese–English McGurk Effects. *Multisensory Research*, *31*(1-2), 79-110. https://doi.org/10.1163/22134808-00002590

Burnham, D., Kasisopa, B., Reid, A., Luksaneeyanawin, S., Lacerda, F., Attina, V., ... & Webster, D. (2015). Universality and language-specific experience in the perception of lexical tone and pitch. *Applied Psycholinguistics*, *36*(6), 1459-1491. https://doi.org/10.1017/S0142716414000496

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*(5572), 1435-1435.

Burnham, D., & Lau, S. (1998). The effect of tonal information on auditory reliance in the McGurk effect. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*. Retrieved from https://www.isca-speech.org/archive_open/avsp98/av98_037.html

Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001). Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers. In AVSP 2001-International Conference on Auditory-Visual Speech Processing, Aalborg, Denmark. Retrieved from https://www.isca-speech.org/archive_open/avsp01/av01_155.html

Burnham, D., Li, W., Carignan, C., Attina, V., Kasisopa, B., & Vatikiotis-Bateson, E. (2019). Visual correlates of Thai lexical tone production: Motion of the head, eyebrows, and larynx?. In *Proceedings of 15th International Conference on Auditory-Visual Speech Processing*. ISCA.

Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., Ciocca, V., Morris, R. H., ... &Jones, C. (2006). The perception and production of phones and tones: The role of rigid and non-rigid face and head motion. Proceedings of the 7th International Seminar on Speech Production, Ubatuba, Brazil. Retrieved from https://ro.uow.edu.au/edupapers/362/

Calvert, G., Spence, C., & Stein, B. (2004). *The handbook of multisensory processes*. Cambridge, Mass: MIT Press.

Campbell, R., Dodd, B., & Burnham, D. (Eds.) (1998). *Hearing by Eye II*. Hove, ES: Psychology Press Ltd..

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 629-647. https://doi.org/10.1002/wcs.63

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, *128*(1), 456-465. https://doi.org/10.1121/1.3445785

Chang, E., Zhou, J., Di, S., Huang, C., & Lee, K. F. (2000). Large vocabulary Mandarin speech recognition with different approaches in modeling tones. In *Sixth International Conference on Spoken Language Processing*. Retrieved from https://www.isca-speech.org/archive/icslp_2000/i00_2983.html

Chao, H., Yang, Z., & Liu, W. (2012, March). Improved tone modeling by exploiting articulatory features for Mandarin speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4741-4744). IEEE. DOI: 10.1109/ICASSP.2012.6288978

Chao, Y. R. (1930*).* A System of Tone Letters. *La Maitre Phonetique 45,* 24-47. Reprinted in *Fangyan 1980.2,* 81-82.

Chen, A. (2003). Reaction time as an indicator to discrete intonational contrasts in English. In *8th European Conference on Speech Communication and Technology* (pp. 97-100). http://hdl.handle.net/11858/00-001M-0000-0013-2790-4

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *The Journal of the Acoustical Society of America*, *123*(4), 2356-2366. https://doi.org/10.1121/1.2839004

Chobert, J., & Besson, M. (2013). Musical expertise and second language learning. *Brain Sciences*, 3(2), 923-940. https://doi.org/10.3390/brainsci3020923

Cienkowski, K. M., & Carney, A. E. (2002). Auditory-visual speech perception and aging. *Ear and hearing*, *23*(5), 439-449.

Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. *The handbook of emotion elicitation and assessment*, *1*(3), 203-221. Retrieved from https://www.pitt.edu/~jeffcohn/biblio/Coan%20013%20chap13.pdf

Cox, R. M., Alexander, G. C., & Gilmore, C. (1987). Intelligibility of average talkers in typical listening environments. *The Journal of the Acoustical Society of America, 81*(5), 1598-1608. https://doi.org/10.1121/1.394512

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106*(2), 633-664. https://doi.org/10.1016/j.cognition.2007.03.013

Cristia, A. (2013). Input to language: The phonetics and perception of infant-directed speech. *Language and Linguistics Compass*, *7*(3), 157-170. https://doi.org/10.1111/lnc3.12015

Cristia, A., & Seidl, A. (2014). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, *41*(4), 913-934. DOI: 10.1017/S0305000914000105

Cutler, A., & Chen, H. C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception & Psychophysics, 59*(2), 165-179.

Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *The Quarterly Journal of Experimental Psychology Section A, 57*(6), 1103-1121. https://doi.org/10.1080/02724980343000701

de Gelder, B., Bertelson, P., Vroomen, J., & Chen, H. C. (1995). Inter-language differences in the McGurk effect for Dutch and Cantonese listeners. In *Fourth*

European Conference on Speech Communication and Technology. Retrieved from https://www.isca-speech.org/archive/archive_papers/eurospeech_1995/e95_1699.pdf

Delogu, F., Lampis, G., & Belardinelli, M. O. (2006). Music-to-language transfer effect: May melodic ability improve learning of tonal languages by native nontonal speakers? *Cognitive Processing 7*(3), 203-207.

Delogu, F., Lampis, G., & Belardinelli, M. O. (2010). From melody to lexical tone: Musical ability enhances specific aspects of foreign language perception. *European Journal of Cognitive Psychology*, *22*(1), 46-61. https://doi.org/10.1080/09541440802708136

Desai, S., Stickney, G., & Zeng, F. G. (2008). Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *The Journal of the Acoustical Society of America*, *123*(1), 428-440. https://doi.org/10.1121/1.2816573

Dörnyei, Z., & Skehan, P. (2003). 18 Individual Differences in Second Language Learning. *The handbook of second language acquisition* (pp.589-623). DOI:10.1002/9780470756492

Ehrman, M. E., Leaver, B. L., & Oxford, R. L. (2003). A brief overview of individual differences in second language learning. *System*, *31*(3), 313-330. https://doi.org/10.1016/S0346-251X(03)00045-9

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ...& Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190-202. DOI: 10.1109/TAFFC.2015.2457417

Ferguson, C. A. (1975). Toward a characterization of English foreigner talk. *Anthropological linguistics*, 1-14. https://www.jstor.org/stable/30027270

Ferguson, C. A. (1981). 'Foreigner talk' as the name of a simplified register. *International Journal of the Sociology of Language, 1981*(28), 9-18. https://doi.org/10.1515/ijsl.1981.28.9

Ferguson, S. H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *The Journal of the Acoustical Society of America, 116*(4), 2365-2373. https://doi.org/10.1121/1.1788730

Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America, 112*(1), 259-271. https://doi.org/10.1121/1.1482078

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research, 50*(5), 1241-1255. https://doi.org/10.1044/1092-4388(2007/087)

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant behavior and development, 10*(3), 279-293. https://doi.org/10.1016/0163-6383(87)90017-8

Fiedler, D. & Müllensiefen, D. (2015). Validation of the Gold-MSI questionnaire to measure musical sophistication of German students at secondary education schools. *Musikpädagogische Forschung / Research in Music Education*, *36*, 199-219. [in German] http://research.gold.ac.uk/id/eprint/17193

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics, 36*(2), 268-294. https://doi.org/10.1016/j.wocn.2007.06.005

Freemake Video Converter (2015). Retrieved from: http://www.freemake.com/

Fu, Q. J., & Zeng, F. G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language and Hearing*, *5*(1), 45-57. https://doi.org/10.1179/136132800807547582

Fuster-Duran, A. (1996). Perception of conflicting audio-visual speech: An examination across Spanish and German. In *Speechreading by humans and machines* (pp. 135-143). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-13015-5_9

Gagné, J. P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in auditory, visual, and audio-visual speech intelligibility for conversational and clear speech. *Journal-Academy of Rehabilitative Audiology, 27*, 135-158.

Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of phonetics*, *11*(2), 149-175. https://doi.org/10.1016/S0095-4470(19)30813-7

Garg, S., Hamarneh, G., Jongman, A., Sereno, J. A., & Wang, Y. (2019). Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. *Speech Communication, 113*, 47-62. https://doi.org/10.1016/j.specom.2019.08.003

Gaser, C., & Schlaug, G. (2003). Brain structures differ between musicians and non-musicians. *Journal of Neuroscience 23*(27), 9240-9245. https://doi.org/10.1523/JNEUROSCI.23-27-09240.2003

Gottfried, T. L., & Riester, D. (2000). Relation of pitch glide perception and Mandarin tone identification. *Journal of the Acoustical Society of America, 108*(5), 2604. Retrieved from https://www2.lawrence.edu/fast/gottfrit/Mandmusic.html

Gottfried, T. L., Staby, A. M., & Ziemer, C. J. (2004). Musical experience and Mandarin tone discrimination and imitation. *Journal of the Acoustical Society of America, 115*(5), 2545. https://doi.org/10.1121/1.4783674

Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, *128*(3), 1322-1336. https://doi.org/10.1121/1.3466853

Grant, K. W., & Braida, L. D. (1991). Evaluating the articulation index for auditory–visual input. *The Journal of the Acoustical Society of America, 89*(6), 2952-2960. https://doi.org/10.1121/1.400733

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104*(4), 2438-2450. https://doi.org/10.1121/1.423751

Grassegger, H. (1995). McGurk effect in German and Hungarian listeners. In *proceedings of the international congress of phonetic sciences, Stockholm* (Vol. 4, No. 3, p. 2).

Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of phonetics*, *32*(3), 395-421. https://doi.org/10.1016/S0095-4470(03)00016-0

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2018). Effects of modality and speaking style on mandarin tone identification by non-native listeners. *Phonetica*, *76*(4), 263-286. https://doi.org/10.1159/000489174

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2019). Mandarin tone identification by tone-naïve musicians and non-musicians in auditory-visual and auditory-only conditions. *Frontiers in Communication*, *4*, 70. https://doi.org/10.3389/fcomm.2019.00070

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2020). Relative Contribution of Auditory and Visual Information to Mandarin Chinese Tone Identification by Native and Tone-naïve Listeners. *Language and speech*, *63*(4), 856-876. https://doi.org/10.1177/0023830919889995

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269-279. https://doi.org/10.1016/j.wocn.2011.11.001

Hassler, M., & Gupta, D. (1993). Functional brain organization, handedness, and immune vulnerability in musicians and non-musicians. *Neuropsychologia*, *31*(7), 655-660. https://doi.org/10.1016/0028-3932(93)90137-O

Hayashi, Y., & Sekiyama, K. (1998). Native-foreign language effect in the McGurk effect: A test with Chinese and Japanese. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*. Retrieved from https://www.isca-speech.org/archive_open/avsp98/av98_061.html

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America, 119*(3), 1740-1751. https://doi.org/10.1121/1.2166611

Helfer, K. S. (1998). Auditory and auditory-visual recognition of clear and conversational speech by older adults. *Journal American Academy of Audiology, 9*, 234-242. Retrieved from https://www.audiology.org/sites/default/files/journal/JAAA_09_03_10.pdf

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research, 53*(2), 298-310. https://doi.org/10.1044/1092-4388(2009/08-0243)

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica*, *33*(5), 353-367. https://doi.org/10.1159/000259792

Jiang, J., Alwan, A., Keating, P. A., Auer, E. T., & Bernstein, L. E. (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Advances in Signal Processing*, *2002*(11), 506945. https://doi.org/10.1155/S1110865702206046

Junqua, J. C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America, 93*(1), 510-524. https://doi.org/10.1121/1.405631

Kalinli, O. (2011). Tone and pitch accent classification using auditory attention cues. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5208-5211). IEEE. doi: 10.1109/ICASSP.2011.5947531.

Kasisopa, B., El-Khoury Antonios, L., Jongman, A., Sereno, J. A., & Burnham, D. (2018). Training children to perceive non-native lexical tones: Tone language

background, bilingualism, and auditory-visual information. *Frontiers in Psychology*, *9*, 1508. https://doi.org/10.3389/fpsyg.2018.01508

Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication*, *57*, 317-330. https://doi.org/10.1016/j.specom.2013.06.003

Kim, J., & Davis, C. (2001). Visible speech cues and auditory detection of spoken sentences: an effect of degree of correlation between acoustic and visual properties. In AVSP 2001-*International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark. Retrieved from https://www.isca-speech.org/archive_open/avsp01/av01_127.html

Koelsch, S., Gunter, T., Friederici, A. D., & Schröger, E. (2000). Brain indices of music processing: "nonmusicians" are musical. *Journal of cognitive neuroscience, 12*(3), 520-541. https://doi.org/10.1162/089892900562183

Kong, Q. M. (1987). Influence of tones upon vowel duration in Cantonese. *Language and Speech, 30*(4), 387-399. https://doi.org/10.1177/002383098703000407

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231-3250. https://doi.org/10.1016/j.pragma.2011.06.008

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396-414. https://doi.org/10.1016/j.jml.2007.06.005

Krause, J. C., & Braida, L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *The Journal of the Acoustical Society of America, 112*(5), 2165-2172. https://doi.org/10.1121/1.1509432

Kricos, P. B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *The Volta Review, 84*(4), 219-225

Krishnan, A., Gandour, J. T., & Bidelman, G. M. (2010). The effects of tone language experience on pitch processing in the brainstem. *Journal of Neurolinguistics*, *23*(1), 81-95. https://doi.org/10.1016/j.jneuroling.2009.09.001

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*(5326), 684-686. DOI: 10.1126/science.277.5326.684

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: continuous or categorical? *Journal of phonetics*, *25*(3), 313-342. https://doi.org/10.1006/jpho.1997.0046

Lee, C. Y., & Hung, T. H. (2008). Identification of Mandarin tones by English-speaking musicians and nonmusicians. *The Journal of the Acoustical Society of America, 124*(5), 3235-3248. https://doi.org/10.1121/1.2990713

Lesner, S. A., & Kricos, P. B. (1981). Visual vowel and diphthong perception across speakers. *Journal of the Academy of Rehabilitative Audiology, 14*, 252-258.

Li, B., & Shuai, L. (2011). *Language Experiences in Perception of Pitches.* Paper presented at 19th Conference of the International Association of Chinese Linguistics, Tianjin, China.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011, March). The computer expression recognition toolbox (CERT). In *Face and gesture 2011* (pp. 298-305). IEEE. doi: 10.1109/FG.2011.5771414.

Liu, S., Del Rio, E., Bradlow, A. R., & Zeng, F. G. (2004). Clear speech perception in acoustic and electric hearing. *The Journal of the Acoustical Society of America, 116*(4), 2374-2383. https://doi.org/10.1121/1.1787528

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America, 94*(3), 1242-1255. https://doi.org/10.1121/1.408177

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y. I., & Yamada, T. (1994). Training Japanese listeners to identify English/r/and/l/. III. Long-term retention of new phonetic categories. *The Journal of the acoustical society of America, 96*(4), 2076-2087. https://doi.org/10.1121/1.410149

Llisterri, J. (1992). Speaking styles in speech research. In *Workshop on Integrating Speech and Natural Language*. Dublin, Ireland. Retrieved from http://liceu.uab.es/~joaquim/publicacions/SpeakingStyles_92.pdf

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/r/and/l: A first report. *The Journal of the Acoustical Society of America, 89*(2), 874-886. https://doi.org/10.1121/1.1894649

Maddieson, I. (2013). *Tone*. In: Dryer, Matthew S. & Haspelmath, Martin (eds.)

Maddieson, I., Dryer, M. S., & Haspelmath, M. (2013). *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Magnotti, J. F., Mallick, D. B., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental brain research*, *233*(9), 2581-2586. https://doi.org/10.1007/s00221-015-4324-7

Maniwa, K., Jongman, A., & Wade, T. (2008). Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *The Journal of the Acoustical Society of America, 123*(2), 1114-1125. https://doi.org/10.1121/1.2821966

Marie, C., Delogu, F., Lampis, G., Belardinelli, M. O., & Besson, M. (2011). Influence of musical expertise on segmental and tonal processing in Mandarin Chinese. *Journal of Cognitive Neuroscience*, *23*(10), 2701-2715. https://doi.org/10.1162/jocn.2010.21585

Marques, C., Moreno, S., Luís Castro, S., & Besson, M. (2007). Musicians detect pitch violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence. *Journal of Cognitive Neuroscience, 19* (9), *1453-1463*. https://doi.org/10.1162/jocn.2007.19.9.1453

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological science*, *26*(3), 341-347. https://doi.org/10.1177/0956797614562453

Massaro, D. W., & Stork, D. G. (1998). Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist*, *86*(3), 236-244. https://www.jstor.org/stable/27857023

Massaro, D.W. (1998). *Perceiving Talking Faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748. https://doi.org/10.1038/264746a0

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing research*, 219(1-2), 36-47. https://doi.org/10.1016/j.heares.2006.05.004

Milovanov, R., Huotilainen, M., Välimäki, V., Esquef, P.A.A., & Tervaniemi, M. (2008). Musical aptitude and second language pronunciation skills in school-aged children: Neural and behavioral evidence. *Brain Research*, *1194*, 81–89. https://doi.org/10.1016/j.brainres.2007.11.042

Milovanov, R., Pietilä, P., Tervaniemi, M., & Esquef, P.A.A. (2010). Foreign language pronunciation skills and musical aptitude: a study of Finnish adults with higher education. *Learning and Individual Differences, 20*(1), 56-60. https://doi.org/10.1016/j.lindif.2009.11.003

Milovanov, R., & Tervaniemi, M. (2011). The interplay between musical and linguistic aptitudes: a review. *Frontiers in psychology*, 2, 321. https://doi.org/10.3389/fpsyg.2011.00321

Mixdorff, H., & Charnvivit, P. (2004). Visual cues in Thai tone recognition. In *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*. Beijing, China. Retrieved from https://www.isca-speech.org/archive/tal2004/tal4_143.html

Mixdorff, H., Charnvivit, P., & Burnham, D. (2005a). Auditory–visual perception of syllabic tones in Thai. In AVSP 2005, *International Conference on Auditory–Visual Speech Processing* (pp. 3–8). Retrieved from https://isca-speech.org/archive_open/archive_papers/avsp05/av05_003.pdf

Mixdorff, H., Hu, Y., & Burnham, D. (2005b). Visual cues in Mandarin tone perception. In *Proceedings of Eurospeech* 2005 (InterSpeech-2005) (pp.405-408). Retrieved from https://www.isca-speech.org/archive/interspeech_2005/i05_0405.html

Mixdorff, H., Lirong, M. C., Nguyen, D. T., & Burnham, D. (2006). Syllabic tone perception in Vietnamese. In *International Symposium on Tonal Aspects of Languages* (pp. 137–142). Retrieved from https://www.isca-speech.org/archive/tal_2006/tal6_091.html

Mok, P. P., & Zuo, D. (2012). The separation between music and speech: evidence from the perception of Cantonese tones. *The Journal of the Acoustical Society of America, 132*(4), 2711-2720. https://doi.org/10.1121/1.4747010

Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., & Besson, M. (2009). Musical training influences linguistic abilities in 8-year-old children: more evidence for brain plasticity. *Cerebral cortex*, *19*(3), 712-723. https://doi.org/10.1093/cercor/bhn120

Müllensiefen, D., Gingras, B., Musil, J., & Stewart L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE*, *9*(2): e89642. https://doi.org/10.1371/journal.pone.0101091

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological research, 71*(1), 4-12. https://doi.org/10.1007/s00426-005-0031-5

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Attention, Perception, & Psychophysics, 57*(7), 989-1001. https://doi.org/10.3758/BF03205458

Ong, J. H., Burnham, D., Escudero, P., & Stevens, C. J. (2017). Effect of linguistic and musical experience on distributional learning of nonnative lexical tones. *Journal of Speech, Language, and Hearing Research, 60*(10), 2769-2780. https://doi.org/10.1044/2016_JSLHR-S-16-0080

Patel, A.D. (2010). *Music, biological evolution, and the brain. In Emerging Disciplines*. Houston, TX, USA: Rice University Press, pp. 91–144.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830. Retrieved from https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, *91*(9), 1306-1326. Doi: 10.1109/JPROC.2003.817150.

Prinzmetal, W., McCool, C., & Park, S. (2005). Attention: reaction time and accuracy reveal different mechanisms. *Journal of Experimental Psychology: General*, *134*(1), 73. https://doi.org/10.1037/0096-3445.134.1.73

Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Retrieved from http://www.pstnet.com.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Rattanasone, N. X., & Best, C. T. (2015). Perceptual assimilation of lexical tone: The roles of language

experience and visual information. *Attention, Perception, & Psychophysics, 77*(2), 571-591. https://doi.org/10.3758/s13414-014-0791-3

Reid, T. B. (1956). Linguistics, structuralism and philology. *Archivum Linguisticum, 8*(1), 28-37.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17*(6), 405-409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D., & Fowler, C. A. (1991). Audiovisual investigation of the loudness-effort effect for speech and nonspeech events. *Journal of Experimental Psychology: Human Perception and Performance, 17*(4), 976. https://doi.org/10.1037/0096-1523.17.4.976

Ryant, N., Yuan, J., & Liberman, M. (2014). Mandarin tone classification without pitch tracking. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4868-4872). IEEE. Doi: 10.1109/ICASSP.2014.6854527.

Schaal, N., Bauer, A.-K., & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrenheit anhand einer deutschen Stichprobe. [The Gold-MSI: Replication and validation of a survey instrument for measuring musical sophistication with a German sample.] *Musicae Scientiae*, *18*(4), 423-447. https://doi.org/10.1177/1029864914541851

Schmidt, R. (2012). Attention, awareness, and individual differences in language learning. *Perspectives on individual characteristics and foreign language education, 6*, 27. Retrieved from https://pdfs.semanticscholar.org/43da/e42c740a9141d8ab4a91222f52d8421e3d4b.pdf

Schneider, K., Dogil, G., & Möbius, B. (2011). Reaction Time and Decision Difficulty in the Perception of Intonation. In *International Speech Communication Association* (INTERSPEECH-2011) (pp. 2221-2224). Retrieved from https://www.isca-speech.org/archive/interspeech_2011/i11_ 2221.html

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology, 41*(3), 341-349. https://doi.org/10.1111/1469-8986.00172.x

Schouten, M. E. (1985). Identification and discrimination of sweep tones. *Perception and Psychophysics, 37*(4), 369–376. https://doi.org/10.3758/BF03211361

Schum, D. J. (1996). Intelligibility of clear and conversational speech of young and elderly talkers. *Journal American academy of audiology, 7,* 212-218. Retrieved from https://www.audiology.org/sites/default/files/journal/JAAA_07_03_10.pdf

Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E), 15*(3), 143-158. https://doi.org/10.1250/ast.15.143

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59*(1), 73-80. https://doi.org/10.3758/BF03206849

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental science, 11*(2), 306-320. https://doi.org/10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., & Tohkura, Y. I. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4), 1797-1805. https://doi.org/10.1121/1.401660

Shaw, J. A., Chen, W.-R., Proctor, M. I., Derrick, D., & Dakhoul, E. (2014). On the inter-dependence of tonal and vocalic production goals in Chinese. Paper presented at the 10th *International Seminar on Speech Production* (ISSP) (pp.395-398). http://hdl.handle.net/10092/10663

Skowronski, M. D., & Harris, J. G. (2006). Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication, 48*(5), 549-558. https://doi.org/10.1016/j.specom.2005.09.003

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *The Journal of the Acoustical Society of America, 118*(3), 1677-1688. https://doi.org/10.1121/1.2000788

Smiljanić, R., & Bradlow, A. R. (2007 August). Clear speech intelligibility: Listener and talker effects. In *the XVIth International Congress of Phonetic Sciences* (pp.661-664). Retrieved from https://www.researchgate.net/profile/Ann_Bradlow/publication/228943878_Clear_speech_intelligibility_listener_and_talker_effects/links/00b4951c9a69228264000000.pdf

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and linguistics compass, 3*(1), 236-264. https://doi.org/10.1111/j.1749-818X.2008.00112.x

Smith, D., & Burnham, D. (2012). Facilitation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, *131*(2), 1480-1489. https://doi.org/10.1121/1.3672703

So, C. K. L. (2006). *Effects of L1 prosodic background and AV training on learning Mandarin tones by speakers of Cantonese, Japanese, and English* (Doctoral dissertation, Department of Linguistics-Simon Fraser University).

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and speech, 53*(2), 273-293. https://doi.org/10.1177/0023830909357156

So, C. K., & Best, C. T. (2011). Categorizing Mandarin tones into listeners' native prosodic categories: The role of phonetic properties. *Poznan Studies in Contemporary Linguistics, 47*(1), 133. https://doi.org/10.2478/psicl-2011-0011

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of/rl/by Japanese adults learning English. *Perception & Psychophysics, 36*(2), 131-145. https://doi.org/10.3758/BF03202673

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661-699. https://doi.org/10.1111/j.0023-8333.2005.00320.x

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the acoustical society of America, 26*(2), 212-215. https://doi.org/10.1121/1.1907309

Swerts, M., & Veldhuis, R. (2001). The effect of speech melody on voice quality. *Speech Communication, 33*(4), 297-303. https://doi.org/10.1016/S0167-6393(00)00061-3

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, *36*(2), 219-238. https://doi.org/10.1016/j.wocn.2007.05.001

Swerts, M., & Krahmer, E. (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics*, *38*(2), 197-206. https://doi.org/10.1016/j.wocn.2009.10.002

Tagarelli, K. M., Ruiz, S., Vega, J. L. M., & Rebuschat, P. (2016). Variability in second language learning: The roles of individual differences, learning conditions, and linguistic complexity. *Studies in Second Language Acquisition, 38*(2), 293-316. doi:10.1017/S0272263116000036

The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/13).

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7*(1), 53-71.

Tillmann, B., Burnham, D., Nguyen, S., Grimault, N., Gosselin, N., & Peretz, I. (2011). Congenital amusia (or tone-deafness) interferes with pitch processing in tone languages. *Frontiers in psychology*, *2*, 120. https://doi.org/10.3389/fpsyg.2011.00120

Tseng, C. Y. (1981). *An acoustic phonetic study on tones in Mandarin Chinese*. Ph.D. dissertation. Providence, RI: Brown University Press.

Tseng, C. Y., Massaro, D. W., & Cohen, M. M. (1986). "Lexical tone perception in Mandarin Chinese: Evaluation and integration of acoustic features". In H. S. R. Kao & R. Hoosain (Eds.), *Linguistics, psychology, and the Chinese language* (pp.91-104). Centre of Asian Studies, University of Hong Kong.

Tupper, P., Leung, K., Wang, Y., Jongman, A., & Sereno, J. A. (2020). Characterizing the distinctive acoustic cues of Mandarin tones. *The Journal of the Acoustical Society of America*, *147*(4), 2570-2580. https://doi.org/10.1121/10.0001024

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and hearing, 28*(5), 656-668. Doi: 10.1097/AUD.0b013e31812f7185

Uchanski, R. M. (2005). Clear speech. In D. B. Pisoni & R. Remez (Eds.), *The handbook of speech perception*. Malden, MA/Oxford, UK: Blackwell.

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech, Language, and Hearing Research, 39*(3), 494-509. https://doi.org/10.1044/jshr.3903.494

Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech communication*, *49*(1), 2-7. https://doi.org/10.1016/j.specom.2006.10.003

Van der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, *60*(4), 448-463. https://doi.org/10.1016/j.jml.2008.12.003

Vanrell, M. D. M., Mascaró, I., Torres-Tamarit, F., & Prieto, P. (2013). Intonation as an encoder of speaker certainty: Information and confirmation yes-no questions in Catalan. *Language and speech, 56*(2), 163-190. https://doi.org/10.1177/0023830912443942

Vatikiotis-Bateson, E., & Yehia, H. (1996). Physiological modeling of facial motion during speech. *Trans. Tech. Comm. Psychol. Physiol. Acoustics*, H-1996-65, 1-8.

Vatikiotis-Bateson, E., Kroos, C., Kuratate, T., Munhall, K. G., & Pitermann, M. (2000). Task constraints on robot realism: The case of talking heads. *9th IEEE International Workshop on Robot and Human Interactive Communication: IEEE RO-MAN 2000 (Cat. No. 00TH8499)* (pp. 352-357). 10.1109/ROMAN.2000.892522

Vatikiotis-Bateson, E., Munhall, K. G., Kasahara, Y., Garcia, F., & Yehia, H. (1996). Characterizing audiovisual information during speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 3, pp. 1485-1488). DOI: 10.1109/ICSLP.1996.607897

Vromans, R., & Postma-Nilsenová, M. (2016). Can musical engagement alleviate age-related decline in inhibitory control? In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Retrieved from https://pure.uvt.nl/ws/portalfiles/portal/18191497/CogSci_VromansPostma2016.pdf

Wang, J., Zhu, Y., Chen, Y., Mamat, A., Yu, M., Zhang, J., & Dang, J. (2020). An eye-tracking study on audiovisual speech perception strategies adopted by normal-hearing and deaf adults under different language familiarities. *Journal of Speech, Language, and Hearing Research*, *63*(7), 2245-2254. https://doi.org/10.1044/2020_JSLHR-19-00223

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*(2), 1033-1043. https://doi.org/10.1121/1.1531176

Wang, Y., Spence, M., Jongman, A., & Sereno, J.A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America, 106*(6), 3649-3659. https://doi.org/10.1121/1.428217

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, *49*(1), 25-47. https://doi.org/10.1159/000261901

Wong, P. C., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*(4), 565. DOI: 10.1017/S0142716407070312

Wong, P. C., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature neuroscience*, 10(4), 420. https://doi.org/10.1038/nn1872

Woodcock, R. W. (1997). *The Woodcock-Johnson tests of cognitive ability—Revised*. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (p. 230–246). The Guilford Press.

Xu, N., & Burnham, D. (2010). Tone hyperarticulation and intonation in Cantonese infant directed speech. In *Speech Prosody 2010-Fifth International Conference*, Chicago, United States. Retrieved from https://www.isca-speech.org/archive/sp2010/sp10_094.html

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America, 111*(3), 1399-1413. https://doi.org/10.1121/1.1445789

Ye, Y., & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes, 14*(5-6), 609-630. https://doi.org/10.1080/016909699386202

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*(3), 555-568. https://doi.org/10.1006/jpho.2002.0165

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

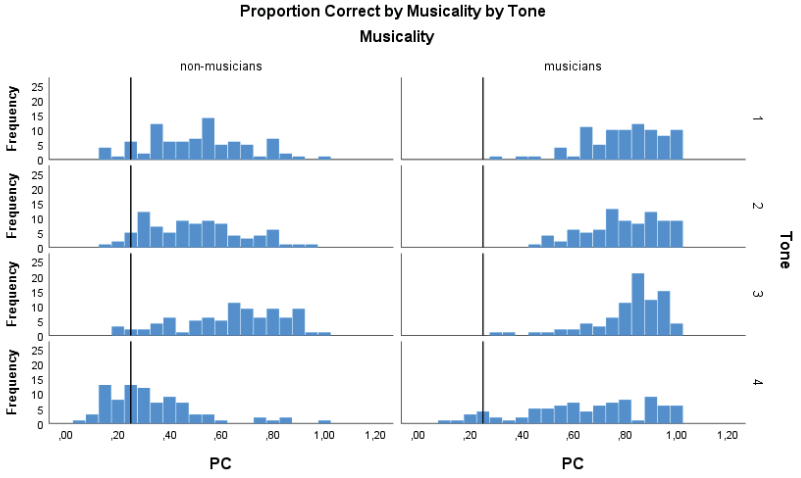
Zheng, A., Hirata, Y., & Kelly, S. D. (2018). Exploring the Effects of Imitating Hand Gestures and Head Nods on L1 and L2 Mandarin Tone Production. *Journal of Speech, Language, and Hearing Research*, 1-17. https://doi.org/10.1044/2018_JSLHR-S-17-0481

Zuccolotto, A.P., Roush, R.E., Eschma, A., & Schneide, W. (2012). E-Prime [Computer software]. Pennsylvania: University of Pittsburgh.

Appendices

mā	má	mǎ	mà
妈	麻	马	骂
yī	yí	yǐ	yì
医	移	椅	意
xiē	xié	xiě	xiè
些	鞋	写	泻
shē	shé	shě	shè
赊	蛇	舍	社
shī	shí	shǐ	shì
师	时	史	市
yōu	yóu	yǒu	yòu
优	由	有	又
fēn	fén	fěn	fèn
分	焚	粉	份
fū	fú	fǔ	fù
夫	浮	斧	妇
pō	pó	pǒ	pò
泼	婆	叵	破
yīng	yíng	yǐng	yìng
鹰	赢	影	硬

Appendix 1. List of words used for producing the stimuli from Chapter 2,3,4,5



Appendix 2. A histogram of each participant's accuracy on individual tone level from Chapter 3. Note that PC represents Proportion Correct. The chance level of the accuracy is .25.

Model 1 (Section 4.3.1)

Accuracy ~ Congruency * Tone + (1|Subject) + (1|Syllable), data = Chinese, family = "binomial"

Information criteria

AIC	*BIC*	*logLikelihood*	*Deviance*	*df residual*
568.8	611.9	-278.4	556.8	9754

Scaled residuals

Min	*1Q*	*Median*	*3Q*	*Max*
-24.2334	0.0354	0.0468	0.0613	0.3627

Random effects

Groups	*Name*	*Variance*	*SD*
Subject	(Intercept)	1.31103	1.1450
Syllable	(Intercept)	0.02904	0.1704

Number of observations: 9760, groups: Subject, 61; Syllable, 10

Fixed effects

	Estimate	*SE*	*z*	*p*
Intercept	4.92530	0.44506	11.067	< 2e-16
Congruency	0.22209	0.75295	0.295	0.76802
Tone	0.48076	0.16267	2.955	0.00312
Congruency*Tone	-0.05291	0.33475	-0.158	0.87440

Correlations of Fixed Effects

	Intercept	Congruency	Tone
Congruency	-0.383		
Tone	-0.707	0.422	
Congruency*Tone	0.346	-0.887	-0.486

Analysis of Variance Table

	df	*Sum of Squares*	*Mean Square*	*F*
Congruency	1	0.1437	0.1437	0.1437
Tone	1	9.4326	9.4326	9.4326
Congruency*Tone	1	0.0217	0.0217	0.0217

Model 2 (Section 4.3.2)

Accuracy ~ Congruency * Tone + (1 + Congruency | Subject) + (1 + Congruency |Syllable), data = Dutch, family = "binomial"

Information criteria

AIC	*BIC*	*logLikelihood*	*Deviance*	*df residual*
16109.1	16183.8	-8044.5	16089.1	12950

Scaled residuals

Min	*1Q*	*Median*	*3Q*	*Max*
-3.7096	-0.7352	-0.5695	1.0554	3.1860

Random effects

Groups	*Name*	*Variance*	*SD*	*Correlation*
Subject	Intercept	0.481539	0.69393	
	Congruency	0.018758	0.13696	0.13
Syllable	Intercept	0.032512	0.18031	
	Congruency	0.005985	0.07736	-0.08

Number of observations: 12960, groups: Subject, 81; Syllable, 10

Fixed effects

	Estimate	*SE*	*z*	*p*
Intercept	-0.35989	0.11001	-3.271	0.00107
Congruency	0.83177	0.11019	7.549	4.39e-14
Tone	-0.09080	0.01984	-4.578	4.70e-06
Congruency*Tone	-0.20838	0.03945	-5.282	1.28e-07

Correlation of Fixed Effects

	Intercept	Congruency	Tone
Congruency	-0.235		
Tone	-0.445	0.445	
Congruency*Tone	0.224	-0.879	-0.503

Analysis of Variance Table

	df	*Sum of Squares*	*Mean Square*	*F*
Congruency	1	38.260	38.260	38.260
Tone	1	70.775	70.775	70.775
Congruency:Tone	1	28.233	28.233	28.233

Model 3 (Section 4.3.3)

Accuracy ~ ConditionAV * Tone + (1 | Subject) + (1 | Syllable), data = AV+AO, family = "binomial")

Information criteria

AIC	*BIC*	*logLikelihood*	*Deviance*	*df residual*
12019.0	12061.7	-6003.5	12007.0	9152

Scaled residuals

Min	*1Q*	*Median*	*3Q*	*Max*
-2.3764	-0.8261	-0.6089	1.0285	2.5614

Random effects

Groups	*Name*	*Variance*	*SD*
Subject	Intercept	0.34840	0.5903
Syllable	Intercept	0.02824	0.1680

Number of observations: 9158, groups: Subject, 118; Syllable, 10

Fixed effects

	Estimate	*SE*	*z*	*p*
Intercept	-0.002796	0.128856	-0.022	0.98269
ConditionAV	0.449349	0.162494	2.765	0.00569
Tone	-0.122704	0.024291	-5.051	4.39e-07
ConditionAV*Tone	-0.173651	0.041692	-4.165	3.11e-05

Correlation of Fixed Effects

	(Intr)	ConditionAV	Tone
ConditionAV	-0.658		
Tone	-0.467	0.370	
CondtionAV*Tone	0.272	-0.631	-0.582

Analysis of Variance Table

	df	*Sum of Squares*	*Mean Square*	*F*
ConditionAV	1	0.046	0.046	0.0462
Tone	1	85.925	85.925	85.9250
ConditionAV:Tone	1	17.587	17.587	17.5868

Appendix 3. Full mixed-effect model outputs for the three models from Chapter 4. All models use *treatment coding* for the independent variables.

Summary

Considering the fact that more than half of the languages spoken in the world (60%-70%) are so-called tone languages (Yip, 2002), and tone is notoriously difficult to learn for westerners, this dissertation focused on tone perception in Mandarin Chinese by tone-naïve speakers. Moreover, it has been shown that speech perception is more than just an auditory phenomenon, especially in situations when the speaker's face is visible. Therefore, the aim of this dissertation is to also study the value of visual information (over and above that of acoustic information) in Mandarin tone perception for tone-naïve perceivers, in combination with other contextual (such as speaking style) and individual factors (such as musical background). Consequently, this dissertation assesses the relative strength of acoustic and visual information in tone perception and tone classification.

In the first two empirical and exploratory studies in **Chapter 2 and 3**, we set out to investigate to what extent tone-naïve perceivers are able to identify Mandarin Chinese tones in isolated words, and whether or not they can benefit from (seeing) the speakers' face, and what the contribution is of a hyperarticulated speaking style, and/or their own musical experience. Respectively, in **Chapter 2** we investigated the effect of visual cues (comparing audio-only with audio-visual presentations) and speaking style (comparing a natural speaking style with a teaching speaking style) on the perception of Mandarin tones by tone-naïve listeners, looking both at the relative strength of these two factors and their possible interactions; **Chapter 3** was concerned with the effects of musicality of the participants (combined with modality) on Mandarin tone perception. In both of these studies, a Mandarin Chinese tone identification experiment was conducted: native speakers of a non-tonal language were asked to distinguish Mandarin Chinese tones based on audio (-only) or video (audio-visual) materials. In order to include variations, the experimental stimuli were recorded using four different speakers in imagined natural and teaching speaking scenarios. The proportion of correct responses (and average reaction times) of the participants were reported.

The tone identification experiment presented in **Chapter 2** showed that the video conditions (audio-visual natural and audio-visual teaching) resulted in an overall higher accuracy in tone perception than the auditory-only conditions (audio-only natural and audio-only teaching), but no better performance was observed in the audio-visual conditions in terms of reaction

time, compared to the auditory-only conditions. Teaching style turned out to make no difference on the speed or accuracy of Mandarin tone perception (as compared to a natural speaking style). Further on, we presented the same experimental materials and procedure in **Chapter 3**, but now with musicians and non-musicians as participants. The Goldsmith Musical Sophistication Index (Gold-MSI) was used to assess the musical aptitude of the participants. The data showed that overall, musicians outperformed non-musicians in the tone identification task in both auditory-visual and auditory-only conditions. Both groups identified tones more accurately in the auditory-visual conditions than in the auditory-only conditions. These results provided further evidence for the view that the availability of visual cues along with auditory information is useful for people who have no knowledge of Mandarin Chinese tones when they need to learn to identify these tones. Out of all the musical skills measured by Gold-MSI, the amount of musical training was the only predictor that had an impact on the accuracy of Mandarin tone perception. These findings suggest that learning to perceive Mandarin tones benefits from musical expertise, and visual information can facilitate Mandarin tone identification, but mainly for tone-naïve non-musicians. In addition, performance differed by tone: musicality improves accuracy for every tone; some tones are easier to identify than others: in particular, the identification of tone 3 (a low-falling-rising) proved to be the easiest, while tone 4 (a high-falling tone) was the most difficult to identify for all participants.

The results of the first two experiments presented in chapters 2 and 3 showed that adding visual cues to clear auditory information facilitated the tone identification for tone-naïve perceivers (there is a significantly higher accuracy in audio-visual condition(s) than in auditory-only condition(s)). This visual facilitation was unaffected by the presence of (hyperarticulated) speaking style or the musical skill of the participants. Moreover, variations in speakers and tones had effects on the accurate identification of Mandarin tones by tone-naïve perceivers.

In **Chapter 4**, we compared the relative contribution of auditory and visual information during Mandarin Chinese tone perception. More specifically, we aimed to answer two questions: firstly, whether or not there is audio-visual integration at the tone level (i.e., we explored perceptual fusion between auditory and visual information). Secondly, we studied how visual information affects tone perception for native speakers and non-native (tone-naïve) speakers. To do this, we constructed various tone combinations of congruent (e.g., an auditory tone 1 paired with a visual tone 1, written as A_xV_x) and incongruent (e.g., an auditory tone 1 paired with a visual tone 2, written as A_xV_y) auditory-visual materials and presented them to native speakers of Mandarin Chinese and speakers of tone-naïve languages. Accuracy, defined as the percentage correct identification of a tone based on its auditory realization, was reported.

When comparing the relative contribution of auditory and visual information during Mandarin Chinese tone perception with congruent and incongruent auditory and visual Chinese material for native speakers of Chinese and non-tonal languages, we found that visual information did not significantly contribute to the tone identification for native speakers of Mandarin Chinese. When there is a discrepancy between visual cues and acoustic information, (native and tone-naïve) participants tend to rely more on the auditory input than on the visual cues. Unlike the native speakers of Mandarin Chinese, tone-naïve participants were significantly influenced by the visual information during their auditory-visual integration, and they identified tones more accurately in congruent stimuli than in incongruent stimuli. In line with our previous work, the tone confusion matrix showed that tone identification varies with individual tones, with tone 3 (the low-dipping tone) being the easiest one to identify, whereas tone 4 (the high-falling tone) was the most difficult one. The results did not show evidence for auditory-visual integration among native participants, while visual information was helpful for tone-naïve participants. However, even for this group, visual information only marginally increased the accuracy in the tone identification task, and this increase depended on the tone in question.

Chapter 5 is another chapter that zooms in on the relative strength of auditory and visual information for tone-naïve perceivers, but from the aspect of tone classification. In this chapter, we studied the acoustic and visual features of the tones produced by native speakers of Mandarin Chinese. Computational models based on acoustic features, visual features and acoustic-visual features were constructed to automatically classify Mandarin tones. Moreover, this study examined what perceivers pick up (perception) from what a speaker does (production, facial expression) by studying both production and perception. To be more specific, this chapter set out to answer: (1) which acoustic and visual features of tones produced by native speakers could be used to automatically classify Mandarin tones. Furthermore, (2) whether or not the features used in tone production are similar to or different from the ones that have cue value for tone-naïve perceivers when they categorize tones; and (3) whether and how visual information (i.e., facial expression and facial pose) contributes to the classification of Mandarin tones over and above the information provided by the acoustic signal. To address these questions, the stimuli that had been recorded (and described in chapter 2) and the response data that had been collected (and reported on in chapter 3) were used. Basic acoustic and visual features were extracted. Based on them, we used Random Forest classification to identify the most important acoustic and visual features for classifying the tones. The classifiers were trained on produced tone classification (given a set of auditory and visual features, predict the produced tone) and on perceived/

responded tone classification (given a set of features, predict the corresponding tone as identified by the participant).

The results showed that acoustic features outperformed visual features for tone classification, both for the classification of the produced and the perceived tone. However, tone-naïve perceivers did revert to the use of visual information in certain cases (when they gave wrong responses). So, visual information does not seem to play a significant role in native speakers' tone production, but tone-naïve perceivers do sometimes consider visual information in their tone identification. These findings provided additional evidence that auditory information is more important than visual information in Mandarin tone perception and tone classification. Notably, visual features contributed to the participants' erroneous performance. This suggests that visual information actually misled tone-naïve perceivers in their task of tone identification. To some extent, this is consistent with our claim that visual cues do influence tone perception. In addition, the ranking of the auditory features and visual features in tone perception showed that the factor perceiver (i.e., the participant) was responsible for the largest amount of variance explained in the responses by our tone-naïve participants, indicating the importance of individual differences in tone perception.

To sum up, perceivers who do not have tone in their language background tend to make use of visual cues from the speakers' faces for their perception of unknown tones (Mandarin Chinese in this dissertation), in addition to the auditory information they clearly also use. However, auditory cues are still the primary source they rely on. There is a consistent finding across the studies that the variations between tones, speakers and participants have an effect on the accuracy of tone identification for tone-naïve speakers.

List of publications

Journal publications

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2018). Effects of modality and speaking style on mandarin tone identification by non-native listeners. *Phonetica*, *76*(4), 263-286. https://doi.org/10.1159/000489174

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2019). Mandarin tone identification by tone-naïve musicians and non-musicians in auditory-visual and auditory-only conditions. *Frontiers in Communication*, *4*, 70. https://doi.org/10.3389/fcomm.2019.00070

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2020). Relative Contribution of Auditory and Visual Information to Mandarin Chinese Tone Identification by Native and Tone-naïve Listeners. *Language and speech*, *63*(4), 856-876. https://doi.org/10.1177/0023830919889995

Han, Y., Castro Ferreira T., Goudbeek, M., Mos, M., & Swerts, M. (Submitted). Automatic Classification of Produced and Perceived Mandarin Tones on the Basis of Acoustic and Visual Properties.

Papers in conference proceedings (peer-reviewed)

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2018, August). Audio-visual Analyses of Differences Between Natural and Teaching Styles for Mandarin Tone Production. In *Proc. 9th International Conference on Speech Prosody 2018* (pp. 729-733). 13-16 June 2018, Poznań, Poland.

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2018, August). Mandarin Tone Identification in Musicians and Non-musicians: Effects of Modality and Speaking Style. In *Proc. TAL2018, Sixth International Symposium on Tonal Aspects of Languages* (pp. 119-123). 18-20 June 2018, Berlin, Germany.

Abstracts of conference presentations (peer-reviewed)

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2016, September). On the effect of modality and speaking style for Mandarin tones classification by L2 Learners. In *the 7th conference on Tone and Intonation in Europe (TIE)*. University of Kent, Canterbury.

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2018 December). Mandarin tone identification by musicians and non-musicians: effects of modality and speaking style. *Dag van de Fonetiek 2018*. Amsterdam, the Netherlands.

Han, Y., Ferreira, T. C., Goudbeek, M., Mos, M., & Swerts, M. (2019, December). Auditory and Visual Cues in the Production and Perception of Mandarin Tones. *Dag van de Fonetiek 2019*. Amsterdam, the Netherlands.

TiCC Ph.D. Series

1. Pashiera Barkhuysen. *Audiovisual Prosody in Interaction.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.
2. Ben Torben-Nielsen. *Dendritic Morphology: Function Shapes Structure.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotor: K.P. Tuyls. Tilburg, 3 December 2008.
3. Hans Stol. *A Framework for Evidence-based Policy Making Using IT.* Promotor: H.J. van den Herik. Tilburg, 21 January 2009.
4. Jeroen Geertzen. *Dialogue Act Recognition and Prediction.* Promotor: H. Bunt. Co-promotor: J.M.B. Terken. Tilburg, 11 February 2009.
5. Sander Canisius. *Structured Prediction for Natural Language Processing.* Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.
6. Fritz Reul. *New Architectures in Computer Chess.* Promotor: H.J. van den Herik. Co-promotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.
7. Laurens van der Maaten. *Feature Extraction from Visual Data.* Promotores: E.O. Postma, H.J. van den Herik. Co-promotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).
8. Stephan Raaijmakers. *Multinomial Language Learning.* Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.
9. Igor Berezhnoy. *Digital Analysis of Paintings.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.
10. Toine Bogers. *Recommender Systems for Social Bookmarking.* Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.
11. Sander Bakkes. *Rapid Adaptation of Video Game AI.* Promotor: H.J. van den Herik. Co-promotor: P. Spronck. Tilburg, 3 March 2010.
12. Maria Mos. *Complex Lexical Items.* Promotor: A.P.J. van den Bosch. Co-promotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).

13. Marieke van Erp. *Accessing Natural History. Discoveries in data cleaning, structuring, and retrieval*. Promotor: A.P.J. van den Bosch. Co-promotor: P.K. Lendvai. Tilburg, 30 June 2010.

14. Edwin Commandeur. *Implicit Causality and Implicit Consequentiality in Language Comprehension*. Promotores: L.G.M. Noordman, W. Vonk. Co-promotor: R. Cozijn. Tilburg, 30 June 2010.

15. Bart Bogaert. *Cloud Content Contention*. Promotores: H.J. van den Herik, E.O. Postma. Tilburg, 30 March 2011.

16. Xiaoyu Mao. *Airport under Control.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotores: N. Roos, A. Salden. Tilburg, 25 May 2011.

17. Olga Petukhova. *Multidimensional Dialogue Modelling*. Promotor: H. Bunt. Tilburg, 1 September 2011.

18. Lisette Mol. *Language in the Hands.* Promotores: E.J. Krahmer, A.A. Maes, M.G.J. Swerts. Tilburg, 7 November 2011 (cum laude).

19. Herman Stehouwer. *Statistical Language Models for Alternative Sequence Selection*. Promotores: A.P.J. van den Bosch, H.J. van den Herik. Co-promotor: M.M. van Zaanen. Tilburg, 7 December 2011.

20. Terry Kakeeto-Aelen. *Relationship Marketing for SMEs in Uganda*. Promotores: J. Chr. van Dalen, H.J. van den Herik. Co-promotor: B.A. Van de Walle. Tilburg, 1 February 2012.

21. Suleman Shahid. *Fun & Face: Exploring Non-Verbal Expressions of Emotion during Playful Interactions*. Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 25 May 2012.

22. Thijs Vis. *Intelligence, Politie en Veiligheidsdienst: Verenigbare Grootheden?* Promotores: T.A. de Roos, H.J. van den Herik, A.C.M. Spapens. Tilburg, 6 June 2012 (in collaboration with the Tilburg School of Law).

23. Nancy Pascall. *Engendering Technology Empowering Women.* Promotores: H.J. van den Herik, M. Diocaretz. Tilburg, 19 November 2012.

24. Agus Gunawan. *Information Access for SMEs in Indonesia*. Promotor: H.J. van den Herik. Co-promotores: M. Wahdan, B.A. Van de Walle. Tilburg, 19 December 2012.

25. Giel van Lankveld. *Quantifying Individual Player Differences.* Promotores: H.J. van den Herik, A.R. Arntz. Co-promotor: P. Spronck. Tilburg, 27 February 2013.

26. Sander Wubben. *Text-to-text Generation Using Monolingual Machine Translation.* Promotores: E.J. Krahmer, A.P.J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.

27. Jeroen Janssens. *Outlier Selection and One-Class Classification*. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 11 June 2013.

28. Martijn Balsters. *Expression and Perception of Emotions: The Case of Depression, Sadness and Fear*. Promotores: E.J. Krahmer, M.G.J. Swerts, A.J.J.M. Vingerhoets. Tilburg, 25 June 2013.

29. Lisanne van Weelden. *Metaphor in Good Shape.* Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 28 June 2013.

30. Ruud Koolen. *"Need I say More? On Overspecification in Definite Reference."* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 20 September 2013.

31. J. Douglas Mastin. *Exploring Infant Engagement. Language Socialization and Vocabulary. Development: A Study of Rural and Urban Communities in Mozambique.* Promotor: A.A. Maes. Co-promotor: P.A. Vogt. Tilburg, 11 October 2013.

32. Philip C. Jackson. Jr. *Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language.* Promotores: H.C. Bunt, W.P.M. Daelemans. Tilburg, 22 April 2014.

33. Jorrig Vogels. *Referential Choices in Language Production: The Role of Accessibility.* Promotores: A.A. Maes, E.J. Krahmer. Tilburg, 23 April 2014 (cum laude).

34. Peter de Kock. *Anticipating Criminal Behaviour.* Promotores: H.J. van den Herik, J.C. Scholtes. Co-promotor: P. Spronck. Tilburg, 10 September 2014.

35. Constantijn Kaland. *Prosodic Marking of Semantic Contrasts: Do Speakers Adapt to Addressees?* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 1 October 2014.

36. Jasmina Marić. *Web Communities, Immigration and Social Capital.* Promotor: H.J. van den Herik. Co-promotores: R. Cozijn, M. Spotti. Tilburg, 18 November 2014.

37. Pauline Meesters. *Intelligent Blauw*. Promotores: H.J. van den Herik, T.A. de Roos. Tilburg, 1 December 2014.

38. Mandy Visser. *Better Use Your Head. How People Learn to Signal Emotions in Social Contexts.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 10 June 2015.

39. Sterling Hutchinson. *How Symbolic and Embodied Representations Work in Concert.* Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2015.

40. Marieke Hoetjes. *Talking hands. Reference in Speech, Gesture and Sign.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 7 October 2015

41. Elisabeth Lubinga. *Stop HIV. Start Talking? The Effects of Rhetorical Figures in Health Messages on Conversations among South African Adolescents*. Promotores: A.A. Maes, C.J.M. Jansen. Tilburg, 16 October 2015.

42. Janet Bagorogoza. *Knowledge Management and High Performance. The Uganda Financial Institutions Models for HPO.* Promotores: H.J. van den Herik, B. van der Walle, Tilburg, 24 November 2015.

43. Hans Westerbeek. *Visual realism: Exploring Effects on Memory, Language Production, Comprehension, and Preference.* Promotores: A.A. Maes, M.G.J. Swerts. Co-promotor: M.A.A. van Amelsvoort. Tilburg, 10 February 2016.

44. Matje van de Camp. *A link to the Past: Constructing Historical Social Networks from Unstructured Data.* Promotores: A.P.J. van den Bosch, E.O. Postma. Tilburg, 2 March 2016.

45. Annemarie Quispel. *Data for all: Data for all: How Professionals and Non-Professionals in Design Use and Evaluate Information Visualizations.* Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 15 June 2016.

46. Rick Tillman. *Language Matters: The Influence of Language and Language Use on Cognition.* Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2016.

47. Ruud Mattheij. *The Eyes Have It*. Promoteres: E.O. Postma, H. J. Van den Herik, and P.H.M. Spronck. Tilburg, 5 October 2016.

48. Marten Pijl. *Tracking of Human Motion over Time.* Promotores: E. H. L. Aarts, M. M. Louwerse. Co-promotor: J. H. M. Korst. Tilburg, 14 December 2016.

49. Yevgen Matusevych. *Learning Constructions from Bilingual Exposure: Computational Studies of Argument Structure Acquisition*. Promotor: A.M. Backus. Co-promotor: A.Alishahi. Tilburg, 19 December 2016.

50. Karin van Nispen. *What Can People with Aphasia Communicate with their Hands? A Study of Representation Techniques in Pantomime and Co-Speech Gesture*. Promotor: E.J. Krahmer. Co-promotor: M. van de Sandt-Koenderman. Tilburg, 19 December 2016.

51. Adriana Baltaretu. *Speaking of Landmarks. How Visual Information Inuences Reference in Spatial Domains.* Promotores: A.A. Maes and E.J. Krahmer. Tilburg, 22 December 2016.

52. Mohamed Abbadi. *Casanova 2, a Domain Specific Language for General Game Development.* Promotores: A.A. Maes, P.H.M. Spronck and A. Cortesi. Co-promotor: G. Maggiore. Tilburg, 10 March 2017.

53. Shoshannah Tekofsky. *You Are Who You Play You Are. Modelling Player Traits from Video Game Behavior.* Promotores: E.O. Postma and P.H.M. Spronck. Tilburg, 19 June 2017.

54. Adel Alhuraibi. *From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT*. Promotores: H.J. van den Herik and Prof. dr. B.A. Van de Walle. Co-promotor: Dr. S. Ankolekar. Tilburg, 26 September 2017.

55. Wilma Latuny. *The Power of Facial Expressions*. Promotores: E.O. Postma and H.J. van den Herik. Tilburg, 29 September 2017.

56. Sylvia Huwaë. *Different Cultures, Different Selves? Suppression of Emotions and Reactions to Transgressions across Cultures*. Promotores: E.J. Krahmer and J. Schaafsma. Tilburg, 11 October, 2017.

57. Mariana Serras Pereira. *A Multimodal Approach to Children's Deceptive Behavior*. Promotor: M. Swerts. Co-promotor: S. Shahid Tilburg, 10 January, 2018.

58. Emmelyn Croes. *Meeting Face-to-Face Online: The Effects of Video-Mediated Communication on Relationship Formation.* Promotores: E.J. Krahmer and M. Antheunis. Co-promotor: A.P. Schouten. Tilburg, 28 March 2018.

59. Lieke van Maastricht. *Second language prosody: Intonation and rhythm in production and perception*. Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 9 May 2018.

60. Nanne van Noord. *Learning visual representations of style*. Promotores: E.O. Postma, M. Louwerse. Tilburg, 16 May 2018.

61. Ingrid Masson Carro. *Handmade: On the cognitive origins of gestural representations*. Promotor: E.J. Krahmer. Co-promotor: M.B. Goudbeek. Tilburg, 25 June 2018.

62. Bart Joosten. *Detecting social signals with spatiotemporal Gabor filters*. Promotores: E.J. Krahmer, E.O. Postma. Tilburg, 29 June 2018

63. Yan Gu. Chinese hands of time: *The effects of language and culture on temporal gestures and spatio-temporal reasoning*. Promotor: M.G.J. Swerts. Co-promotores: M.W. Hoetjes, R. Cozijn. Tilburg, 5 June 2018.

64. Thiago Castro Ferreira. *Advances in natural language generation: Generating varied outputs from semantic inputs*. Promotor: E.J. Krahmer. Co-promotor: S. Wubben. Tilburg, 19 September 2018.

65. Yu Gu. *Automatic emotion recognition from Mandarin speech*. Promotores: E.O. Postma, H.J. van den Herik, H.X. Lin. Tilburg, 28 November 2018.

66. Francesco Di Giacomo. *Metacasanova: A high-performance meta-compiler for domain-specific languages*. Promotores: P.H.M Spronck, A. Cortesi, E.O. Postma. Tilburg, 19 November 2018.

67. Ákos Kádár. *Learning visually grounded and multilingual representations*. Promotores: E.O. Postma, A. Alishahi. Co-promotor: G.A. Chrupala. Tilburg, 13 November 2019.

68. Phoebe Mui. *The many faces of smiling: Social and cultural factors in the display and perception of smiles*. Promotor: M.G.J. Swerts. Co-promotor: M.B. Goudbeek. Tilburg, 18 December 2019.

69. Véronique Verhagen. *Illuminating variation: Individual differences in entrenchment of multi-word units*. Promotor: A.M. Backus. Co-promotores: M.B.J. Mos, J. Schilperoord. Tilburg, 10 January 2020 (cum laude).

70. Debby Damen. *Taking perspective in communication: Exploring what it takes to change perspectives*. Promotor: E.J. Krahmer. Co-promotores: M.A.A. Van Amelsvoort, P.J. Van der Wijst. Tilburg, 4 November 2020.

71. Alain Hong. *Women in the Lead: Gender, Leadership Emergence, and Negotiation Behavior from a Social Role Perspective*. Promotor: J. Schaafsma. Co-promotor: P.J. van der Wijst. Tilburg, 3 June 2020.

72. Chrissy Cook. *Everything You Never Wanted to Know about Trolls: An Interdisciplinary Exploration of the Who's, What's and Why's of Trolling in Online Games*. Promotores: J. Schaafsma, M.L. Antheunis. Tilburg, 22 January 2021.

73. Nadine Braun. *Affective Words and the Company They Keep: Investigating the interplay of emotion and language*. Promotor: E.J. Krahmer. Co-promotor: M.B. Goudbeek. Tilburg, 29 March 2021.

74. Yueqiao Han. *Chinese Tones: Can You Listen with Your Eyes? The Influence of Visual Information on Auditory Perception of Chinese Tones*. Promotor: M.G.J. Swerts. Co-promoter: M.B.J. Mos, M.B. Goudbeek. Tilburg, 18 June 2021.