LUNDS UNIVERSITET

Naïve listeners' perceptual learning of Chinese tones:

The influence of L1 background in prosody

and the effect of an auditory-image (AI) training paradigm

Lili Wen

Supervisor: Mikael Roll, Michael Schoenhals (co-supervisor)

Centre for Language and Literature, Lund University

MA in Language and Linguistics, Chinese

Degree Project – Master's Thesis, 30 Credits

September 2018

Abstract

The present perceptual learning study was conducted in an attempt to investigate to which extent naïve listeners' L1 prosodic background affected their learning performance of a lexical tone language and whether an auditory-image (AI) training paradigm curtailed the online-learning process of lexical tones. Thirty-three Mandarin-naïve Swedish-native listeners (NS listeners) and 33 Mandarin-naïve English-native listeners (NE listeners) participated in a behaviour experiment designed to answer the question of whether native listeners of a pitch accent language (NS listeners) had an advantage over native listeners of a non-tonal non-pitch-accent language (NE listeners) in learning to perceive tones in Mandarin Chinese (MC). The results obtained – by on-line measuring the two groups' general mean values of response accuracy (RA) and response time (RT) for the whole experiment and by comparing the groups' block-to-block improvements in terms of mean values of RA and RT in both all and matched trials – indicated that (a) both NS and NE listeners made significant improvements across blocks in identifying tones in MC; (b) NS listeners were quicker learners of tones in MC compared with NE listeners; (c) NS listeners outperformed NE listeners in both matched (significantly) and mismatched (non-significantly) trials; (d) NS listeners' intrinsic pre-activation of suffixes associated with initial stem tone to some extent impeded the online-acquiring of some tone combinations in MC. Furthermore, the AI training paradigm is more effective compared to traditional auditory training paradigm in terms of percentage gain (in correct identification) and training time. The results suggest that the short-term laboratory perceptual learning of lexical tones at a naïve level is determined by factors such as linguistic experience, training paradigm and procedure, and influenced by factors such as musical aptitude, psychoacoustic ground for and physiological bias toward perception of certain pitch patterns.

Different possible explanations were discussed regarding why initial falling tone combinations with high initial stem tone in MC were processed slower and less accurately compared with initial falling tone combinations with low initial stem tone for NS listeners. It remains an open question whether NS listeners' intrinsic use of stem tone in predicting the incoming suffixes is transferable in perceiving other non-native tones aside from MC, until more studies are carried out concerning NS listeners' perceptual learning on other tonal or pitch accent languages at same phonological level.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Mikael Roll, who suggested the research method for this thesis. Without his great generosity with time, his outstanding expertise, and his patience and tolerance with my mistakes, difficulties and setbacks, this psycholinguistic study could never have been possible for me with a study area of Chinese in Sweden.

I would like to thank Professor Michael Schoenhals for his kind and professional arrangement of an exchange term financed with a scholarship from the Birgit Rausing Language Program.

In addition, I would also like to thank Stefan Lindgren and Joost van de Weijer, for their support at the Humanities Laboratory, Lund University; Mechtild Tronnier, for the knowledge and inspirations from many of her courses in phonetics and precious suggestions for the amending of this thesis; Ioanna Dimitrakopoulou, for the perspectives and feedback.

And last, but not least, I would like to thank all my participants and all those who helped me in some way during this unbelievably long and crooked journey. Without their participation and their help, this thesis could not be possible either.

Table of Contents

List of figures vii
List of tables ix
Abbreviations x
Preface xi
1. Introduction 1
1.1 Motivation for the Study 1
1.2 Main research question and hypothesis of the study 1
1.3 Other concerns of the study 4
1.3.1 Swedish as a studying language in perception of Chinese tones 4
1.3.2 Mandarin naïve listeners as participants 9
1.3.3 Disyllabic tokens as sound stimuli and systematic learning of a tonal system 10
1.3.4 An AI perceptual training paradigm 11
1.4 The organization of the study 13
2. Background 14
2.1 Physical, auditory and physiological correlates of tone 14
2.2 The three languages of the study 15
2.2.1 Mandarin Chinese 15
2.2.2 Swedish 23
2.2.3 English 27
2.2.4 Comparisons among MC, Swedish and English 28
2.3 Top-down and bottom-up processing in speech word recognition 31
2.4 General influencing factors in pitch perception 33
2.4.1 Psychoacoustic ground for pitch perception 33
2.4.2 Music-to-language transfer 35
2.5 Perceptual learning of lexical tones 36
2.5.1 The effects of perceptual training 36
2.5.2 Issue of ID and SD training procedure in tone perceptual training 39
2.6 The present study 40
2.6.1 RQ 1: The influence of L1 prosodic background 41
2.6.2 RQ 2: The efficacy of AI training paradigm 42
2.6.3 Other factors of note 43
3. Method 44
3.1 Participants 44

3.2 Stimuli and design 45
3.2.1 Sound stimuli 45
3.2.2 Image stimuli 47
3.2.3 Design list of stimuli 48
3.2.4 Experimental design 49
3.3 Procedure 49
3.4 Data analysis 51
3.4.1 Response accuracy (RA) 51
3.4.2 Response time (RT) 54
4. Results 55
4.1 Overall performance 55
4.1.1 Total scores for each group for the whole experiment 55
4.1.2 Effects of sound and image match/mismatch on learning performance 56
4.2 Block-by-block improvements on perceptual learning 58
4.2.1 Improvements for both groups and for each group across blocks 58
4.2.2 Summary of improvements of RA and RT for all and matched trials across blocks 71
4.3 Effects of L1 background on learning performance 73
4.3.1 Effects of L1 background on learning performance for matched trials 74
4.3.2 Effects of L1 background on learning performance for mismatched trials 78
4.3.3 Summary of effects of L1 background on learning performance for matched and mismatched trials 82
5. Discussion 86
5.1 Effects of L1 background on learning performance 86
5.1.1 Initial and final tone sub-condition 86
5.1.2 General learning performance in terms of the easiest and most confusing tone combinations 93
5.2 Effects of AI training paradigm on learning performance 94
5.3 Effects of sound and image match/mismatch on learning performance 103
5.4 Influence of musical aptitude on learning performance 104
5.4.1 Influence of musical aptitude on improvements of RA and RT for all and matched trials across blocks 104
5.4.2 Influence of musical aptitude on initial and final sub-condition 104
5.5 Other factors of concern 106
6. Conclusion 108
7. References 111

8. Appendix 118
8.1 Appendix 1 118
8.2 Appendix 2 119
8.3 Appendix 3 122
8.4 Appendix 4 125
8.5 Appendix 5 128

List of figures

Figure 2.1 F0 patterns in MC on the syllable *ma* in MC……..16

Figure 2.2 Amplitude contours on the syllable *ma* in MC……20

Figure 2.3 Intensity contours on the syllable *ma* in MC……...21

Figure 2.4 F0 patterns (non-focal) in Central Swedish on the disyllabic word *anden*…….....25

Figure 2.5 F0 patterns (focal) in Central Swedish on the disyllabic word *anden*…… 26

Figure 2.6 Stress patterns on the word *SUBject* and *subJECT* in English……28

Figure 2.7 The F0 contour of T3 (falling part) and T4 on the syllable *pa* in MC…….29

Figure 2.8 Differential threshold for frequency as influenced by stimulus duration……34

Figure 2.9 Human audibility – pitch and energy distribution in speech and music…….36

Figure 3.1 The waveforms and pitches of the four tone combinations with *pa4* [p^ha^{51}] as the first syllable……...47

Figure 4.1 Mean scores of RA (a) and RT (b) for matched and mismatched trials for NS and NE listeners……...57

Figure 4.2 Mean values of RA for all trials across blocks for both language groups……...59

Figure 4.3 Mean values of RA of 16 tone combinations of NS listeners across blocks for all trials……...61

Figure 4.4 Mean values of RA of 16 tone combinations of NE listeners across blocks for all trials……...62

Figure 4.5 Mean values of RA of 16 tone combinations of NS and NE listeners across blocks for all trials……62

Figure 4.6 Mean values of RT for all trials across blocks for both language groups……...63

Figure 4.7 Mean values of RA for matched trials across blocks for both language groups…….....65

Figure 4.8 Mean values of RA of 16 tone combinations of NS listeners across blocks for matched trials……68

Figure 4.9 Mean values of RA of 16 tone combinations of NE listeners across blocks for matched trials……68

Figure 4.10 Mean values of RA of 16 tone combinations of NS and NE listeners across blocks for matched trials……..69

Figure 4.11 Mean values of RT for matched trials across blocks for both language groups……70

Figure 4.12 Mean values of RA of initial (a) and final (b) tone sub-condition for matched trials for both language groups……..75

Figure 4.13 Mean values of RT of initial (a) and final (b) tone sub-condition for matched trials for both language groups……...78

Figure 4.14 Mean values of RA of initial (a) and final (b) tone sub-condition for mismatched trials for both language groups……..80

Figure 4.15 Mean values of RT of initial (a) and final (b) tone sub-condition for mismatched trials for both language groups……..81

List of tables

Table 2.1 The pitch value (Hz) and duration (ms) of the four tones in MC on the syllable *ma* ……22

Table 2.2 The duration, pitch range and ΔF/duration of the four pitch falls…….....30

Table 4.1 Estimated marginal means and effect of L1 background with four set ups of variates for both language groups……...56

Table 4.2 Response accuracy in percentage across blocks for all trials for each language group ……59

Table 4.3 Improvements in percentage of correct response between blocks for all trials for each language group…….59

Table 4.4 Pairwise comparisons of estimated marginal means of RA for all trials for NS and NE listeners……...59

Table 4.5 Pairwise comparisons of estimated marginal means of RT for all trials for NS and NE listeners……...64

Table 4.6 Response accuracy in percentage across blocks for matched trials for each language group……65

Table 4.7 Improvements in percentage of correct response between blocks for matched trials for each language group……65

Table 4.8 Pairwise comparisons of estimated marginal means of RA for matched trials for NS and NE listeners……67

Table 4.9 Pairwise comparisons of estimated marginal means of RT for matched trials for NS and NE listeners……71

Abbreviations

MC: Mandarin Chinese

NS: Native Swedish

NE: Native English

AI: Auditory-image

RA: Response accuracy

RT: Response time

NJ: Native Japanese

NMC: Native Mandarin Chinese

ITI: Inter-trial-interval

ID: Identification

SD: Discrimination

Preface

It is of course true in a purely physical sense that all tone languages can be analysed in terms of register. All this means is that you can't have "movement" without having an A which you move to B. Hence, we may claim necessary priority of points A and B to movement from A to B. But movement from A to B (a physical statement) is not really the same thing as movement in a direction away from A and toward B. Psychologically, one can have an experience of such movement without being able to "geometrize" in terms of fixed points A and B.

– Edward T. Sapir (Pike 1948:8)

1. Introduction

1.1 Motivation for the Study

Native linguistic experience (L1) forms our perception of sounds in other languages (L2) (Best, 1995). Extensive research effort has been devoted to cross-linguistic study on how L1 backgrounds in prosody influence L2 perception and learning of lexical tones (Gandour, 1983; Kaan, Barkley, Bao, & Wayland, 2008; Francis, Ciocca, Ma & Fenn, 2008; Wayland & Li, 2008; Burnham et al., 2015). Among these studies, to our knowledge, few has employed Swedish as a learning language in such research contexts. To fill this gap, the current perceptual learning study of lexical tones in Modern Chinese employs native speakers of both Swedish and English as participants. The motivation for the study is twofold. On the one hand, this study aims to explore whether listeners of a native pitch accent language have an advantage over listeners of native non-lexical-tone and non-pitch-accent language in identifying and differentiating tones from a lexical tone language with which they have no prior experience. And on the other, this study attempts to verify whether adopting an auditory-image (AI) training paradigm will be more effective on listeners' learning performance in perceiving Mandarin tones by comparing previous similar studies in which traditional auditory training paradigm was adopted (at the results and discussion part).

1.2 Main research question and hypothesis of the study

Mandarin Chinese[1] (henceforth, MC) is a lexical tone language with four distinctive phonological tones: high even (˥), middle rising (˧˥), low falling-rising (˨˩˧) and high falling (˥˩).

[1] Mandarin Chinese refers to **Standard Chinese**. **Standard Chinese**, also known as **Modern Standard Chinese**, **Modern Standard Mandarin**, **Standard Mandarin**, or simply **Mandarin**, is a standard variety of Chinese with its pronunciation based on Beijing dialect and its grammar based on written vernacular Chinese. The term Mandarin Chinese is referred in this study for the sake of simplicity and continuity with previous similar studies.

An example of this is the syllable *ma* that can mean “mother”, “hemp”, “horse” and “curse” depending on which tone it refers to.

Swedish is conventionally categorized as a pitch accent language or a word accent language. There are pitch differences between accent 1 and accent 2 which are associated with the stressed syllable and conditioned by inflectional or derivation suffixes (Bruce, 1977; Roll, Horne & Lindgren, 2010; Söderström, Roll & Horne, 2012). An example of this is the word *anden*, which, with a low stem tone (*and-en*: word stem *and* [ˈaɬnd] + definitive suffix *-en*) means “the duck,” and with a high stem tone (*ande-n*: word stem *ande* [ˋaɬndə] + definitive suffix-*n*) means “the spirit.”

Both Swedish word accents and Chinese lexical tones share the trait of using pitch differences to distinguish the meaning of words. However, in MC each syllable carries one tone and syllables composed of identical segments but carrying different tones differ in meaning while in Swedish accent 1 and accent 2 do not contrast at a monosyllabic level. In terms of psycholinguistic models, Chinese tones are assumed to be an obligatory part of the phonological representation of words in the mental lexicon whereas Swedish word accents are rather termed as “morphophonological” in which a low stem tone is a default accent 1 tone and high accent 2 tone is induced onto word stems by particular suffixes rather than being an integrated part of a particular word in the mental lexicon as in Chinese (Rischel, 1963; Riad, 1998; Roll et al., 2010; Roll, Söderström & Horne, 2013; Roll, Söderström, Mannfolk, Shtyrov, Johansson, Westen & Horne, 2015; Roll, Söderström, Frid, Mannfolk & Horne, 2017; Ingram 2007:26).

Thus, Swedish (to a limited extent, more details see 2.2.2.1) and Chinese share the characteristic of using pitch variations to distinguish lexical meanings of words. However, the function of pitch variations differs. If the main function of tones in monosyllabic words in MC is disambiguation, the function of pitch accents in Swedish is rather anticipatory (Roll, et

al., 2015; Söderström, Horne, Frid & Roll, 2016), besides the intrinsic tone perception coding system used by their native speakers differs as well. Nonetheless, to date, little is known about whether native experience of a pitch accent language such as Swedish gives advantage over native experience of a nontonal language such as English in perceiving a lexical tone language such as MC with which one has no prior experience. On the other hand, a rather large number of cross-linguistic studies examined non-native lexical tone perception by both tonal (tonal here referring lexical tone) and nontonal language speakers, both behaviourally (Wayland & Guion 2004; Wayland et al., 2008; Hao, 2012; Li, Shao & Bao, 2017) and neurolinguistically (Kaan et al., 2008; Xu, Gandour, Talavage, Wong, Dzemidzic, Tong et al., 2006), and many of such studies found that prior experience with a lexical tone language such as MC was beneficial in discriminating (Kaan, et al., 2008; Burnham et al., 2015), discriminating and identifying (Wayland et al., 2004), discriminating or identifying (Wayland, et al., 2008) tone contrasts in another lexical tone language such as Thai, or vice versa, native experience of Thai contributed to the identification (Li, 2016) or discrimination (Li, et al., 2017) of non-native MC tones compared to native experience of a nontonal language. The present study attempted to bridge the gap by employing a psycholinguistic behaviour experiment designed to answer the question of whether native speakers of a pitch accent language outperform native speakers of a nontonal language in identifying and differentiating tones from a lexical tone language. In addition, by using an untraditional training procedure, the present study investigates the step-by-step improvements of the two language groups in perceiving tones in MC that they are naïve to at the beginning of the training.

Bearing in mind that a pitch accent language lies between a lexical tone language and a nontonal language in terms of phonological category, it seems reasonable to hypothesize that native Swedish speakers' native pitch accents could both facilitate and/or interfere at the same

time[2] in naïve perception of tones in MC though at different dimensions, considering the similarities and differences between the two languages.

1.3 Other concerns of the study

Before going further into the organization of the present study, it is worth mentioning that this study differs from the previous similar studies in the following aspects.

1.3.1 Swedish as a studying language in perception of Chinese tones

Mother tongue's influence on perception of Chinese tones has been well investigated (Gandour & Harshman 1977; Gandour 1978, 1983; Kiriloff, 1969; Leather, 1990; Shen, 1991; So, 2006; So & Best 2010; Wang, Spence, Jongman, & Sereno, 1999; Wang, Jongman, & Sereno, 2003; Wang, 2013; White, 1981). By and large, as mentioned in 1.2, a variety of cross-linguistic behavioural and neuroimaging research found that linguistic experience of lexical tone language (Thai or MC) facilitates non-native lexical tone perception (MC or Thai) (Kaan, et al., 2008; Li, 2016; Li, et al., 2017; Wayland et al., 2004; Wayland et al., 2008; Schaefer & Darcy, 2014). However, among this research, some behavioural studies found that linguistic experience of a tonal language (Hmong, Cantonese) did not necessarily facilitate non-native tone perception (MC) (Wang 2013; So & Best, 2010; Li et al., 2017; Li & Zhang, 2010). Thus, tone language speakers' experience with lexical tones in general may not always give them advantage over the perception of non-native lexical tones.

The argument for the mixed results lies in that tonal languages differ in how they make use of phonetic features (e.g. pitch height and pitch contour) for tonal contrasts (Li et al., 2017), and native tonal language speakers process non-native tones with reference to their native phonological tone categories (Wang, 2013). Lexical tones in Thai (Thai has 4 phonemic tones and one "common" tone) and MC (MC has four phonemic tones) are comparable with each

[2]It may seem ambiguous that two incompatible conjectures are included in one hypothesis. However, according to previous cross-linguistic studies concerning perceptual learning of lexical tones, linguistic experience of a lexical tone language or a pitch accent language both facilitated (in terms of general performance) and hindered (in terms of tone confusions) the learning of a non-native lexical tone language (So & Best 2010; Li et al., 2017).

other in terms of the number of tone categories and the pitch patterns of the tones. Therefore, it is not difficult for native Thai speakers to map the tones in MC onto their own and likewise, native MC speakers to map the tones in Thai onto their own[3], but such corresponding mapping in tonal system is missing between Hmong (Hmong has seven phonemic tones) and MC. In the case of Cantonese (Cantonese has six phonemic tones), high even tone and high falling tone are allotones in Cantonese but belong to two different tonemes (T1 and T4) in MC and thus cause further confusion in perception of MC for native Cantonese speakers (So et al., 2010). In sum, if both L1 and L2 are tonal, then perceptual accuracy in L2 depends on how phonetic features and phonemic status of tone contrasts vary between L1 and L2.

From the perspective of cross-linguistic tone coding system, Xu, et al. (2006) investigated lexical tone processing by using functional magnetic resonance imaging (fMRI). Both Thai and Chinese tones were superposed on Chinese syllables and used as stimuli. They found overlapping activation for prelexical processing and lexical processing of tones in the temporal plane of the left superior temporal gyrus (STG) in both native Thai and MC participants. The authors thus concluded that the left STG is sensitive to language experience in terms of lexical tones and plays a critical role in phonological decoding.

In a related study, by using both event-related potentials (ERPs) and fMRI, Roll et al. (2015) found left STG to be activated for native Swedish speakers in predicting the incoming endings of words when hearing the stem tone at the beginning of words. Thus, left STG seemed

[3] T1 (Thai) = T3 sandhi (MC), T2 (Thai) = T4 (MC), T3 (Thai) ≈ T2 (MC) and T4 (Thai) = T3 (MC). Native Thai L2 learners of MC usually adopt this mapping strategy in learning tones in MC and acquire (both quickly and effortlessly) their own version of tone normalization pattern in MC that is differentiable for native speakers of MC (Chen, 2007).

important even for native Swedish speakers to decode their native phonological tones although they were not directly associated with lexical phonemes as in Thai and MC.

Like Swedish, Japanese also uses pitch variations to distinguish meanings at word level and the pitch patterns do not contrast at monosyllabic level either in Japanese. The accent in Japanese is realised in the mora that carries a high pitch, and the pitches of the first and second mora must differ (HL or LH) in disyllabic words (Gussenhoven, 2004; Dong, Tsubota & Dantsuji, 2013; So et al., 2010). According to Shibata & Shibata (1990), the percentage of minimal pairs that are strictly distinguished by pitch variations in Japanese lies below 10% which is comparable with Swedish in terms of the number of minimal pairs distinguished by pitch variations. Thus, both being categorized as pitch accent languages, Swedish and Japanese are comparable in terms of the number of minimal pairs and the patterns of pitch variation.

So et al. (2010) investigated perceptual learning of four tones in monosyllabic tokens in MC by Mandarin-naïve Hong Kong Cantonese, Japanese and Canadian English listeners. The results showed that native Japanese listeners made significantly fewer tonal errors in the identification tasks than other language groups and the Japanese group had the best learning results compared to the other two language groups. Employing the Perceptual Assimilation Model (PAM), the authors interpreted the results as T2 (mid rising) and T4 (high falling) in MC were assimilated to the Japanese Low-High (LH) and High-Low (HL) pitch accent patterns respectively, thus Japanese pitch accent system (HL & LH) facilitates the perception and learning of tones in Mandarin. However, the results of the study also found that both the English and the Japanese listeners showed more confusions in T1 (high even) and T2 (middle rising) pair than the Cantonese listeners and this confusion could hardly be solely explained by the PAM model.

Based on the results from So et al. (2010), it seems that pitch accents in Japanese facilitate the perception and the learning of tones in MC. It would thus be significant to investigate whether some tones in MC can also be assimilated into pitch accents in Swedish as well and if or how Japanese and Swedish differ in perceiving and learning of tones in MC with regard to the similarities in terms of the prosodic system between these two languages.

Although Japanese and Swedish share the similarities in terms of the number of minimal pairs and the patterns of pitch variation, however, as mentioned in 1.3, Swedish has stress differences that can be described in much the same way as stress differences in English and its pitch differences between accent 1 and accent 2 are only contrasted on the stressed syllable while Japanese is in some ways between a tone language and a stress language and the accent in Japanese is invariably realized in high pitch and the pitch pattern of a word is predictable if the position of the word accent is known (Tsujimura 2006: 74). Thus, the predictive function of word accents differs between Swedish and Japanese. Unlike word accents in Swedish which are used in morphophonological prediction (Roll et al., 2010, 2013, 2015), the pre-head anticipation (Japanese is a head-final language, head referring to a verb) function of pitch accents in Japanese was found not necessarily active for native Japanese speakers when accessing the concept of a lexical item in the processing of accent-contrasted homophonic pairs in simple sentences in Tamaoka, Saito, Kiyama, Timmer & Verdonschot's electrophysiological study (2014). It seems that more investigations need to be conducted concerning the predictive function of word accents in Japanese for making such a conclusion. Nevertheless, based on the studies insofar, Japanese and Swedish share the similarity of using pitch variations to distinguish lexical meanings at word level but differ in their function as anticipation of the incoming word suffixes (due to the regularity in Swedish of pitch variations in relation to word inflection). It would thus be engrossing to see if using the pitch variations in the stem tone of the stressed vowel as anticipatory linguistic cue in Swedish will

be transferred as perceptual cue in perception of tones in MC by employing native Swedish speakers as participants in the present behavioural study.

In addition, another relevant language that has garnered relatively less attention in this cross-linguistic perception of lexical tone research context is Korean. It has been reported that Korean South Kyengsang dialect employs a pitch accent prosodic system similar to Tokyo Japanese in that accented moras bear a high tone and unaccented moras bear a default low tone (Narahara, 1985). However, in Standard Korean pitch variations are not lexically contrastive. In Schaefer & Darcy's study (2014) of cross-linguistic (Mandarin Chinese, Japanese, English, Standard Korean and Korean South Kyengsang dialect) perception of Thai tones, it was found that, in terms of the overall credits, the linguistic experience of a tonal language (MC) outperformed all other groups, followed by the linguistic experience of a pitch accent language (Japanese and Korean South Kyengsang dialect), and by the linguistic experience of a non-lexical-tone and non-pitch-accent language (English and Korean Seoul dialect). It seems that the pitch accents in Korean South Kyengsang dialect benefits the perception of tones in Thai compared to Standard Korean in which the characteristics of pitch accent is missing.

With a speech perceptual perspective, Chang (2012) asserts that there are two tonal patterns in South Kyengsang dialect: a high tone with higher initial F0, an early timing of the F0 peak and a shorter syllable duration, and a rising tone with lower initial F0, a later timing of the F0 peak and a longer syllable duration. In his study of perception of South Kyengsang Korean tones, timing of F0 peak, initial F0 and syllable duration were reported as the crucial acoustic parameters for native perception of South Kyengsang dialect (Chang, 2012). Wang & Li (2011) investigated perceptions of T2 (middle rising) and T3 (low falling-rising) of MC by native Korean speakers and found that native Korean learners of MC at beginners' level perceived T3 in MC as a middle rising tone instead of a falling-rising tone when the initial F0

was low due to the low initial F0 was associated with a rising tone in Korean. The perception result of this study might also be interpreted as the initial F0 having been transferred as acoustic cue in non-native perception of lexical tones in MC by native Korean learners of MC at beginners' level (negative transfer of linguistic experience in L1). As mentioned earlier in this introduction part, the low pitch height of the stem tone in stressed syllable is associated with the default accent 1 in Swedish (Roll et al., 2015), it would thus be interesting to investigate whether native speakers of Swedish will use the low stem tone in stressed syllable as predictive acoustic cue in perception of tones in MC with T3 at the initial syllable. If the transfer takes place for native speakers of Swedish, it would be expected to be a positive transfer of linguistic experience in L1 because both the initial low tone in Swedish accent 1 and T3 in MC anticipate an initial pitch fall.

1.3.2 Mandarin naïve listeners as participants

Focused on identification of monosyllabic tokens, Gao (2016) studied the perception of tones in MC by native Swedish speakers and found that T3 (a low falling tone) was the easiest tone to identify, followed by T4 (a high falling tone). This result is in line with the hypothesis that the low stem tone in accent 1 words is assigned as a default word accent in Swedish (Riad, 1998, 2009) and with the ERPs study of Swedish word processing in Roll, et al. (2010). Based on these two aforementioned studies (Gao, 2016 & Roll et al., 2010), it is tempting to presume that native Swedish speakers at least partly perceive words in MC with T3 at initial syllable as accent 1 (a default tone) in Swedish and thus the easiest tone, and words with T4 at the initial syllable as accent 2 in Swedish and thus next easiest tone. However, Gao employed native Swedish speakers with more than one year's learning experience of MC as participants. In order to have a more precise understanding of the influence of a L1 prosodic system on adult L2 learners' perceptual performance of tones in MC with the least interference of earlier lexical tone experience and thus to establish a baseline for acquisition of a tonal L2, the focus of this study is on Mandarin-naïve (= non-learner of Mandarin) native Swedish listeners

(henceforth, NS listeners) and Mandarin-naïve native English listeners (henceforth, NE listeners) are employed as control group. Hence the influence of L2 knowledge and learning conditions on listeners' learning performance in the present study is excluded.

In a cross-linguistic study of perception of Thai tones, Burnham et al. (1996) reported that native speakers of Swedish demonstrated comparable scores in terms of mean value of accuracy rates and reaction times to those of Cantonese speakers in a tone discrimination task. In another study, Burnham et al. (2015) investigated naïve perception of Thai tones by native speakers of lexical tone languages (Thai, Cantonese, and Mandarin), a pitch accent language (Swedish), and a nontonal language (English) and found that native speakers of lexical tone and pitch accent language were better than native speakers of nontonal language in auditory-only and auditory-visual discrimination of Thai tones. It would thus be worth to further investigate whether native Swedish speakers would be better than native English speakers in naïve perceiving tones in MC as well. Furthermore, pitch ranges in MC were reported as wider than Japanese, Cantonese, Thai and English (Tsukada, Kondo & Sunaoka, 2016; So & Best, 2010; Burnham et al., 2015; Chen, 1974; White, 1981). In other words, pitch ranges of tones in Thai are narrower compare to pitch ranges of tones in MC. Thus, pitch ranges of tones in Swedish and Thai are presumed to be more similar than pitch ranges of tones in Swedish and MC. This aspect further makes Swedish an interesting language to investigate in the present study. The differences in pitch ranges of the languages in concern will be discussed in the next part "language background" of this study.

1.3.3 Disyllabic tokens as sound stimuli and systematic learning of a tonal system

In their study concerning online speech processing of pitch movements signalling lexical (present in tonal and pitch accent languages only) and postlexical (present in all languages) contrasts by both Mandarin and Dutch speakers, Braun & Johnson (2011) suggested that pitch contours that broadly mirror intonational patterns in L1 might be perceived with more success

when the stimulus is bisyllabic or multisyllabic word (or pseudoword). Likewise, Lee & Nusbaum (1993) recommended that a bisyllabic or multisyllabic word (or pseudoword) makes it possible for speakers of non-lexical tone languages to make tonal distinctions for specific F0 patterns. Take MC in concern, the majority of words in MC is disyllabic. The disyllabic words make up 69.8% of the total words (Duanmu, 2007:160). In Swedish, accent 2 is only seen in words with more than one syllable (Elert, 1981). Moreover, pitch variations occur on individual syllables in lexical tone languages whereas in pitch accent languages in general, it is the pitch variations between successively syllables that is essential. Thus, the influence of L1 background in prosody will be most properly studied if the training stimuli are based at the same phonological level. There are four tones at the monosyllabic level in MC and disyllabic tokens will constitute totally 16 possible tone combinations. Thus, a systematic learning of a tonal system in MC will be carried out in this study by employing 16 disyllabic tone combinations of the same token as sound stimuli.

1.3.4 An AI perceptual training paradigm

Several previous studies on L2 perception of tones adopted traditional auditory training paradigm (Chen, 2012; Wang et al., 1999; Wang, 2013) and the training results differed among the studies by the same training paradigm. However, pitch differences signal different levels of prosodic contrasts in all languages, only when they are used to distinguish lexical meanings are categorised as lexical pitch accent or lexical tone languages. Just as McCawley cited in Fromkin (1978) put it: “What is basic to the role of pitch in a tone language is not its contrastiveness but its lexicalness” (p. 3). Thus, one important component of tone perceptual training is to set different tone patterns in a lexical context. In Chandrasekaran, Sampath and Wong’s (2010) lexical tone (four Mandarin tones) learning studies by Mandarin-naïve native English speakers, a sound-to-meaning training paradigm was employed and proved to benefit the learning. Wong and Perrachione (2007) trained Mandarin-naïve native English-speaking adults to learn to use pitch patterns to identify a vocabulary of six English pseudosyllables

superimposed with three pitch patterns resembling three Mandarin tones and found that a phonetic–phonological–lexical continuity could improve phonological awareness and general auditory ability in adult non-native tonal word learning.

Thus, in order to increase tone-meaning awareness in an even broader way, this study employs a training procedure of a comparably better ecological validity in the sense that each tone combination to be trained in the experiment will be associated with an animal. The participants' task is to determine which tone combination belongs to which animal and which does not, in a way not unlike how children hear and learn their first language when using picture books[4] or in a way that is comparable to second-language word learning in a sense that it gradually builds up a mental mapping between a new non-native pitch pattern and a known concept during the training process. Thus, the online acquisition of different pitch patterns is put in a linguistic context.

In sum, the present study tries to answer the question of whether linguistic experience of a pitch accent language have any facilitatory effects on learning to perceive tones in a lexical tone language compared to linguistic experience of a nontonal language. It is hypothesized that native experience of a pitch accent language can both facilitate and hinder the learning of tones in a lexical tone language. The aim of the study is, theoretically, to investigate the influence of learners' L1 prosodic system, and methodologically, to examine the efficacy of an AI training paradigm – on the performance of the learners' naïve perception of lexical tones.

[4] The auditory-image approach works quite like children's picture books ("pekböcker" in Swedish): The adults point to a picture of a monkey (just as an example) and tell the child: "This is a monkey!" After some sweet explanations, the adults point to the same picture again and ask the child: "Is this a monkey?" Now and then the adults pick up a picture of a bear and ask the child if it is a monkey or not just to make sure that the child has learned. Similar way will be adopted in this study.

1.4 The organization of the study

The present study consists of six parts: introduction, background, method, results, discussion and conclusion. In the introduction, the statement of purpose of the study, the main research question and hypothesis, a brief review of previous studies concerning influence of L1 background in prosody on perception of lexical tones that led to the research question and other aspects in which this study differs from the previous similar studies were included.

The language background, the theoretical background of tone perception, and AI training paradigm in perceptual learning of lexical tones build up the main framework of the background of the thesis. Based on this framework, the research questions, hypotheses and predictions are raised again and unfolded further at the final section of this part.

The design list of the stimuli, the experimental design, the procedure of the experiment and data analysis are described in the method. In the results part, the effects of L1 background on learning performance in both matched and mismatched trials and in both initial and final tone sub-condition are reported and compared and the effect of AI training paradigm on perceptual learning is reported by measuring block-by-block improvements.

In the fifth part, effects of L1 background in both initial and final tone sub-condition, AI training paradigm, and sound and image match/mismatch on perceptual learning are generalized and discussed. The answers to research questions and the validity of hypotheses and predictions, and the direction of further research are presented and summarized in the conclusion part.

2. Background

This background part begins with a short introduction of physical, auditory and physiological correlates of tone, followed by descriptions of the three studying languages. In the description of tones in MC, perceptual cues such as F0, amplitude patterns, duration and phonation mode are unfolded and discussed, and in the description of pitch accents in Swedish, previous studies of online processing of Swedish word accents are reviewed. Later, top-down and bottom-up processing in speech word recognition, general influencing factors such as psychoacoustic ground for pitch perception and music-to-language transfer in pitch perception are presented. Furthermore, previous studies are reviewed and compared regarding perceptual learning of lexical tones with traditional auditory training paradigm and sound-to-meaning training paradigm, and with identification and/or discrimination as training procedure, and hence different aspects of the AI training paradigm in this study are raised. Finally, the research questions, hypotheses and predictions are presented in more detail under the heading of "the present study".

2.1 Physical, auditory and physiological correlates of tone

The vocal folds vibrate to open and close the airflow in quick succession and thus sounds are produced. The rate of the vibration is called fundamental frequency (F0) (Duanmu, 2007:225). F0 is a physical (acoustic) term and it denotes how many cycles per second a sound signal contains. It is measured in Hertz (Hz). For example, a F0 of 250 Hz means 250 cycles per second (Yip, 2002:289). Pitch, on the other hand, is generally considered as the perceived height of F0 or an auditory impression of F0 (Ladefoged, 2003:75). Pitch is thus an auditory term or a perceptual term. Pitch can be a property of speech signals as well as non-speech signals such as music (Yip, 2002:289). In music, pitches are selected from fixed hierarchical scales (e.g. scalar). In contrast to music, pitch in language (e.g. contour tones in tonal languages) is not solely organized in such a hierarchical scale, it is rather continuous and

curvilinear in nature (Bildelman, Gandour & Krishnan, 2011). Thus, there is no "in-" or "out-of-tone" in languages, but rather "native" or "non-native" tones.

When pitch variations affect the meaning of a word, they are called tones (Ladefoged et al., 2011:255). Tone is thus a linguistic term that refers to a phonological category that distinguishes phonemes (or words). When tone is described or perceived, F0 as well as amplitude (or intensity, perceived as loudness) and duration are usually considered as its most critical physical correlates. However, tone cannot only be determined by the dimension of the rate of flow between vocal folds and the pressure drop across vocal folds, but also by the dimension of adjustments of the laryngeal muscles often referred as phonation (ladefoged, 2003:86; Ladefoged et al., 2011:254; Yang, 2011). Thus, phonation is a physiological correlate of tone and pitch is an auditory correlate of tone. In some tonal languages in the world variations in pitch occur in association with a certain 'phonation mode'. For example, the initial part of T3 in MC is a falling pitch with a low pitch onset and is thus often associated with creaky phonation.

2.2. The three languages of the study

2.2.1 Mandarin Chinese

2.2.1.1 General description

As mentioned earlier in this study, Mandarin Chinese (MC) is a typical lexical tone language in that syllables composed of identical segments but carrying different tones (i.e., different pitch patterns) differ in lexical meaning. MC has four distinctive tones in its phonological inventory.

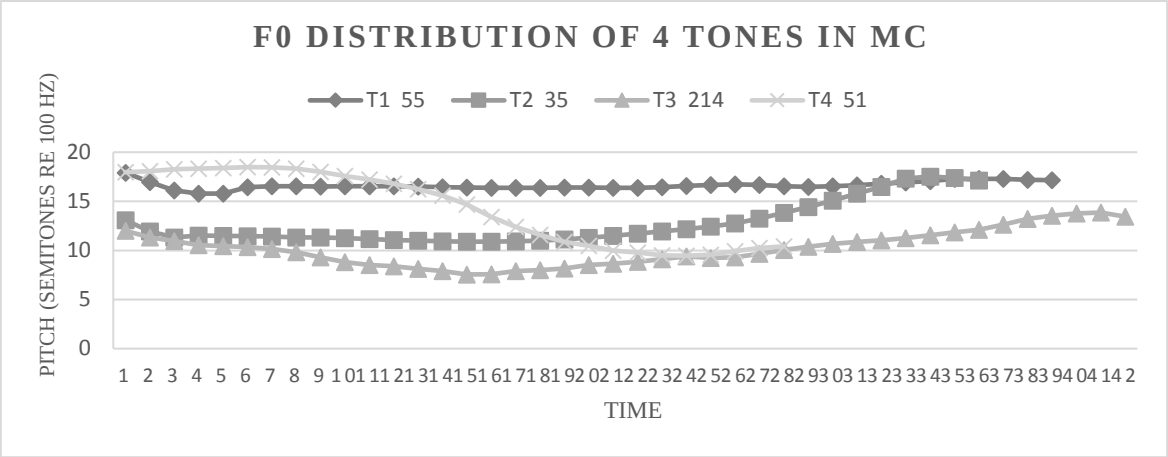


Figure 2.1 F0 patterns in MC on the syllable *ma* spoken by a female native speaker of MC. The pitch points were tracked every 10 milliseconds.

As illustrated in Figure 2.1, Tone 1[5] (T1), also named *yīnpíng* 阴平 in MC, is a high level tone with pitch sustained constantly high at pitch level 5[6]. Tone 2 (T2), also named *yángpíng* 阳平 in MC, is a middle rising tone with the pitch increasing from level 3 to level 5. Tone 3 (T3), also named *shǎngshēng* 上声 in MC, is a falling-rising tone with the pitch first dips from level 2 to level 1 and then rises to level 4. Tone 4 (T4), also named *qùshēng* 去声 in MC, is a "full fall that starts from level 5 and glides all the way down to level 1" (Kong & Zhang, 2017).

[5] Tone type such as "T1 or tone 1" is the most common tone marking way in academic journals outside the Chinese speaking world, bearing in mind that tone types such as "T1 or tone 1" that are conventionally used in specifying tones in MC do not have the same pitch levels (or contours) as those conventionally used for Thai tones. For example, "T1 or tone 1" in MC is a high even tone (˥) whereas "T1 or tone 1" is a low falling tone (˩) in Thai (Ladefoged et al., 2011:258). The Chinese names for the four tones such as *shǎngshēng* 上声 and *qùshēng* 去声 originated from Middle Chinese (latest) according to the earliest preserved book about Chinese prosody *qièyùn*《切韵》.T1 and T2 belong to the same tone termed as level "*píng* 平", in contrast to the other three tones that collectively referred as oblique "zè 仄" in Middle Chinese. Due to historic splits and mergers, none of the modern varieties of Chinese have the exact four tones as in Middle Chinese. (Tang, 2013:2; Guo, Wang & Lu, 2012).

[6] This tone marking system is known as "Five degrees standard transferring method" (also be named as Chao's digitals 五度标调法) in which 5 represents the highest pitch and 1 the lowest pitch in an individual's pitch range.

In nature speech, T3 goes through T3 Sandhi[7] in MC. In this study T3 in all stimuli were in their citation forms. The exclusion of tone 3 sandhi rule was to facilitate the overall learnability of the AI training paradigm that was designed for naïve listeners of MC.

2.2.1.2 Cues in tone perception of MC

2.2.1.2.1 F0

As mentioned earlier in 2.1, F0 is assumed to be the primary cue for pitch, and thus, tone perception. After having investigated the previous research on tone perception, Gandour (1978) summarized that F0 is the absolutely necessary cue to tone perception in a wide range of seemingly unrelated tonal languages such as MC, Thai, Yoruba and Swedish. Although this finding has been widely referred in cross-linguistic tone perception studies, it is noteworthy that different pitch patterns imposed on the stimuli in his study were obtained from speakers of different L1 backgrounds instead of native speakers of one certain tonal language and thus the results of the study was criticized as "not best representative for the actual tone perception" (Zsiga & Nitisaroj, 2007; Shen, 2016). However, the four tones in MC do not only differ in the F0 value of pitch onset and offset, but also in duration, amplitude pattern and voice quality. Thus far, how these different dimensions of the four tones in MC contribute to its perception (primary or secondary cues) might be a question of relevance in this study. The following paragraphs will give a short information of these relevant perceptual cues.

2.2.1.2.2 Pitch height and pitch contour

Among the four tones in MC, T1 and T4 have the similar starting point of pitch height, and T1 and T2 have the similar endpoint of pitch height. So et al. (2010) found that tone pairs in MC that shared similar phonetic figures such as similar pitch onset (T1 and T4) or offset (T1 and T2) were more confusing than tone pairs with dissimilar phonetic figures (T1 and T3) for non-native perception. Moreover, for native tone perception of MC, the order of tone pairs

[7] In tone 3 Sandhi in MC, a T3 becomes a T2 when it precedes another T3 whereas a T3 loses its "rising tail" and becomes a low-dip tone [21] when it precedes T1, T2 and T4.

could influence the perception of the high even tone (T1) in synthetic stimuli with the pitch onset and offset of the second tone (high falling tone, T4) manipulated. Lin & Wang (1985) found that as the pitch onset of the second tone (high falling tone, T4) increased, the chance of the first high even tone (T1) perceived as middle rising tone (T2) also increased. In a pilot study about categorical perception between T1 and T4 in MC with the offset of the second tone (T4) manipulated in stimuli, it was found that the first high even tone (T1) tended to be perceived as middle rising tone (T2) as the pitch range of the second tone decreased.

Other studies found that non-native speakers tended to pay more attention to the onset and offset and the average pitch in perception of tones in MC while native speakers assigned more perceptual weights on pitch contour than pitch height (Gandour, 1983; Chandrasekaran, Krishnan & Gandour, 2007). The results explain why native speakers perceive tones in a categorical manner only between level tone and contour tone or between contour tones, but never between level tones.

2.2.1.2.3 Amplitude

Whalen & Xu (1992) found that there was a close correlation between amplitude contour and pitch contour in MC. For example, T4 was short and T3 was long both on amplitude contour and pitch contour, the amplitude contour of T4 showed the similar steep slope as the pitch contour of T4. By using stimuli that both formant and F0 information were completely missing and duration was controlled, they found that the identification rate of T1, T2 and T4 was 55.3%, 69,5% and 92.3% respectively by native MC speakers with only the amplitude contour as perceptual cue.

Seen in the sound stimuli used at the practice part of the experiment in this thesis, the amplitude contour of T4 and the pitch contour of T4 on the syllable *ma* has similar falling rate. However, one thing is worth noticing, the amplitude contours of T1 and T2 are both

falling and look similar for the syllable *ma* whereas the pitch patterns of T1 and T2 for the syllable *ma* differ (illustrated in Figure 2.2 and 2.1).

From the perspective of physiological mechanism of tone production, T1 in MC is produced with the cricothyroid muscle[8] contracted. The contraction of the cricothyroid muscle causes the increasing longitudinal tension of vocal folds and they become thinner and their opening become consequently narrower on account of their being lengthened (Sonesson, 1968:63). When the adductive tension between the vocal folds is kept stable, the pitch of T1 is sustainable high. On the other hand, with rising tone frequency (T2 in MC), besides the cricothyroid muscle contracts to increase longitudinal tension of vocal folds, the vocalis muscle (the medial portion of the thyroarytenoid muscle) alters the tension vs. relaxation of the vocal folds and the timing and tension of the contraction of the vocalis muscle regulate the degree and velocity of pitch rising[9]. At the time when the opening of the vocal folds is restricted, the pressure of the flow of air out of the lungs increases and causes the increasing of vibration rate of the vocal folds and thus also causes the increase in F0. This increased pressure of the air flow in producing T2 in MC at the comparably early stage of the time course of the syllable makes the amplitude patterns of T1 and T2 on the syllable *ma* look similar. The amplitude patterns of the four tones on the syllable *ma* in MC is illustrated in Figure 2.2.

[8] For the sake of simplicity, the muscles here are mentioned in the singular though the corresponding muscles of the two sides under normal conditions always work together.

[9] The process of increasing pitch is rather a complicated process, besides the contraction of the cricothyroid muscle increases the tension of the vocal fold and the vocalis muscle regulates the tension of the vocal folds, the contraction of other larynx muscles is also involved in the process. For example, the contraction of the lateral cricoarytenoid muscle, the transverse arytenoid muscle and the oblique arytenoid muscle leads to the opening of the rima glottis, the contraction of thyroepiglottic muscle decreases the diameter of the aditus larynges. Furthermore, the contraction of the sternothyroid muscle rocks the thyroid forward and down and thus assists in increasing the tension of the vocal folds as well.

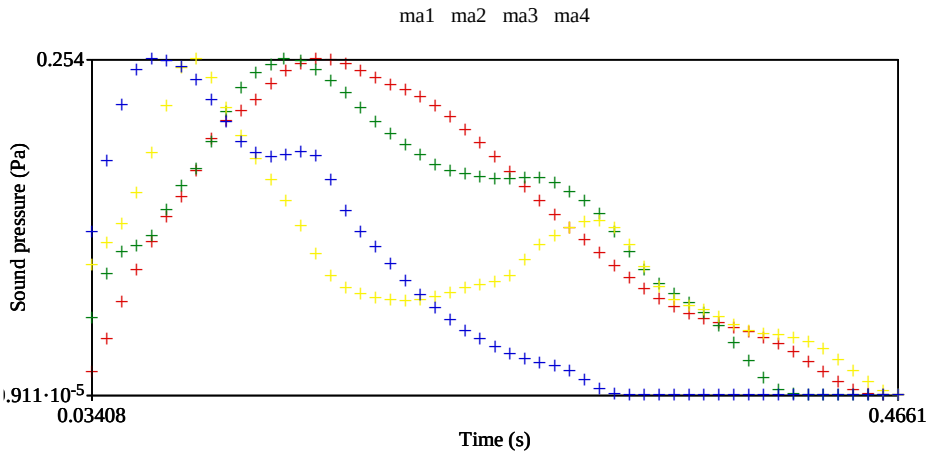


Figure 2.2 Amplitude contours on the syllable *ma* in MC spoken by a female native speaker of MC (ma1_red ma2_green ma3_yellow ma4_blue).

It is necessary to point out that the sound stimuli were manipulated in such a way that both the F0 and formant were completely lost and syllable carrier *yi* (stimulus) had a glide as its onset in Whalen et. al. (1992) whereas in the present study natural stimuli are used and the syllable carrier has a nasal as its initial, which explains why the amplitude contours of T1 and T2 mirror each other in this study but not in Whalen et al. (1992).

Since "intensity is dependent on the amplitude of the sound wave, the size of the vibration in air pressure" (Ladefoged, 2003:90), the intensity patterns of the four tones resemble the amplitude patterns of the four tones on the syllable *ma* in MC in the present study. The intensity patterns of the four tones in MC are illustrated in Figure 2.3. In English stress differences occur between syllables of words, intensity (perceived as loudness) is considered as one of the indicators of stress, thus T1 and T2 could possibly be misperceived as equivalent in terms of pitch instead of loudness for native speakers of languages with stress differences but not lexical pitch variances at word level.

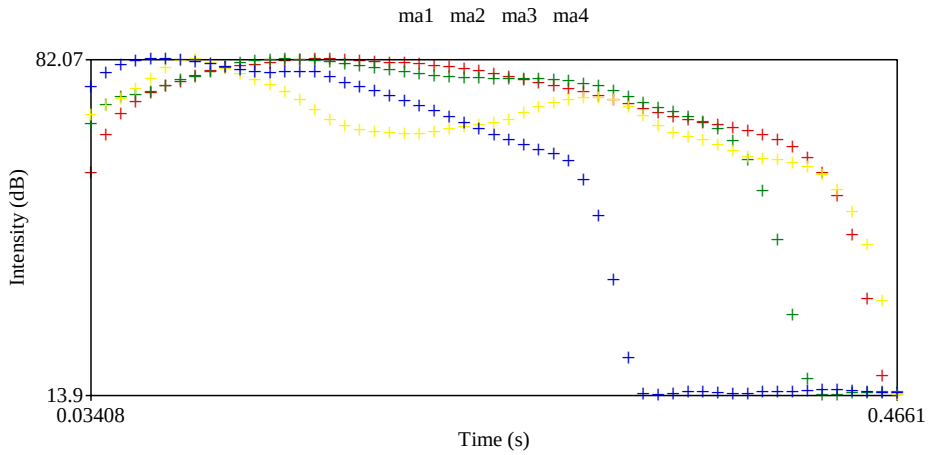


Figure 2.3 Intensity contours on the syllable *ma* in MC spoken by a female native speaker of MC (ma1_red ma2_green ma3_yellow ma4_blue).

2.2.1.2.4 Voice quality

As for the dimension of voice quality (also termed as phonation mode) of the four tones in MC, T3 is often produced with creaky phonation while T1, T2 and T4 are produced with modal phonation (Duanmu.2007:229). Gårding et al. (1986) found that creaky phonation was a concomitant but not a necessary cue in native perception of T3 and T4 in a series of manipulated pitch movements over the syllable of *mai* from T4 to T3. Li & Xu (2018) investigated native perception of whispered Mandarin and found that tones were much worse identified in whispered phonation compared to modal phonation (49.9% and 96.4% respectively), but T3 was much better identified compared to all the other three tones in MC with an identification rate as high as 84.57% in whispered phonation. Even when F0 was removed from phonated stimuli in amplitude-modulated noise, T3 was still best judged compared to other tones in MC. By using a pitch synchronous overlap-add (PSOLA) method in producing the speech stimuli and applying an inverse- filtering method in assessing the invariability of sound source of the stimuli, Yang (2011) found that phonation factors (creaky voice) were adopted in native perception of tones in MC. Thus, it is tempting to assume that creaky phonation is used as one of the perception cues in native tone perception in MC.

2.2.1.2.5 Duration

Previous studies (Liu &Samuel, 2004; Whalen & Xu, 1992) have assumed that tones in MC tend to have some intrinsic durational differences: T3 tends to be the longest and T4 to be the shortest. The pitch value and duration of the four tones in MC on the syllable *ma* are illustrated in table 2.1.

Table 2.1 The pitch value (Hz) and duration (ms) of the four tones in MC on the syllable *ma* spoken by a female native MC speaker.

/ma/	*Durations on the vowel part (ms)*	*F0 (Hz)*		
		Pitch onset	*Turning point*	*Pitch offset*
Tone 1	391	259		267
Tone 2	363	229		271
Tone 3	421	183	154 (at 145ms)	222
Tone 4	270	282		172

Whalen et al. (1992) and Liang (1963) found that there was no tendency for the tone to be correctly perceived based only on duration for native MC perceivers while other studies had different results. By manipulating the number of noise bands with the spectral detail within each band removed, Fu, Zeng, Shannon & Soli (1998) found that identification accuracy for the four tones in MC in native perception was at about 80% in general regardless the number of bands, among which the correct identification rate for T3 and T4 was higher than T1 and T2. The authors thus concluded that temporal waveform envelop cues (duration) played an important role in native perception of at least T3 and T4. Kong & Zeng (2006) found that with temporal envelop (duration) cues, native speakers of MC could achieve tone identification accuracy at 70%–80% correct level in quiet, but the performance was lower in noise.

In sum, there are different dimensions of the four tones in MC that contribute to their perception. However, F0 in pitch height has been the primary cue in most previous studies. Furthermore, some studies have reported that amplitude information alone can be used as primary cue in native tone identification in MC when F0 information is missing. Liu &Samuel (2004), Li & Xu (2018) and Yang (2011) found that phonation factors (creaky voice) were adopted in native perception of tones in MC when F0 information were neutralized. A few

studies have reported that temporal envelop (duration) can be used as the only perceptual cue in native identification of tones in MC when information from F0 and spectral envelop (amplitude) are extracted from the stimuli.

2.2.2 Swedish

2.2.2.1 Pitch accent system in Swedish

Unlike MC, which is categorized as a lexical tone language, Swedish is rather categorized as lexical pitch accent languages or word accent languages along with Japanese, Norwegian, Basque ect. There is no general consensus on the exact typological distinction between lexical tones and pitch accents in speech prosody, for the reason that "[it seems difficult to draw a dividing line between languages with contrastive tone on (almost) all syllables and languages with tone contrasts in more restricted locations in the world"] (Gussenhoven, 2004:47).

Swedish has a binary lexical and phonological tone contrast on the syllable with the primary stress and distinctive tones are referred as accent 1 and accent 2. The distinction is typically illustrated with minimal pairs such as *anden* with accent 1 it means "the duck" and with accent 2 it means "the spirit" (Riad, 2014:182). However, seen from the details of the decomposed morphological constructions of the minimal pairs per se, these two words have a slightly different morphological construction [with the accent 1 forms typically being monosyllabic stems (e.g. *and* – 'duck'), and accent 2 typically being represented by disyllabic stems (e.g. *ande* – 'spirit') (Riad, 2014:182)]. There are about 350 minimal tone pairs relying on pitch accent contrast in Swedish (Elert, 1971:19).

There are different regional varieties of Swedish with different phonetic accent patterns, and a variety of Swedish spoken in Finland does not have the word tone accents at all (Bruce & Gårding, 1978). However, only Central Swedish variety is considered for the present study.

2.2.2.2 Non-focal accent 1 and 2 in Central Swedish

In Central Swedish, there is different tonal timing of an underlying HL between accent 1 and accent 2 in relation to the beginning of the stressed syllable (Bruce, 1977). For non-focal accent 1, the high tone can be first present at the pretonic syllable (the syllable that precedes the stressed syllable) and then falls to a low tone at the beginning of the stressed vowel (L*). The high tone will not be realised if the non-focal accent 1 word begins with a stressed vowel, which begins directly with a low stem tone (L*). For non-focal accent 2, the high tone is realized at the beginning of the stressed vowel and then falls to a low tone in parallel with the time course of the stressed syllable (Söderström et al, 2012).

According to previous studies (Bruce,1977; Söderström et al., 2012), non-focal accent 1 falls from an F0-maximum in the pretonic syllable to an F0-minimum at the end of the stressed vowel (for the word beginning with vowel, non-focal accent 1 falls from an F0-maximum to an F0-minimum over the time course of the stressed vowel) while non-focal accent 2 falls from an F0-maximum in the stressed vowel to an F0-minimum in the post-stress vowel. Thus, the duration of the fall for non-focal accent 1 is shorter than non-focal accent 2. The frequency ranges from F0-maximum to F0-minimum for non-focal accent 1 and non-focal accent 2 was as small as negligible in Bruce (1977).

The gap between F0-maximum and F0-minimum for non-focal accent 1 and non-focal accent 2 is presumed to be similar. However, the timing of the peak of F0 contour differs between non-focal accent 1 and non-focal accent 2 in association with the stressed syllable. For non-focal accent 1 the peak appears as early as the pretonic syllable (often in the preceding word). For non-focal accent 2, on the other hand, the peak starts at the beginning of stressed syllable. Thus, the peak of F0 contour appears earlier for non-focal accent 1 than for non-focal accent 2. Considering that the fall for non-focal accent 1 is shorter in duration than non-focal accent 2, the gradient for the fall for non-focal accent 1 is thus steeper than for non-focal accent 2

(Bruce 1977). F0 patterns (non-focal) in Central Swedish on the disyllabic word *anden* are illustrated in figure 2.4 below.

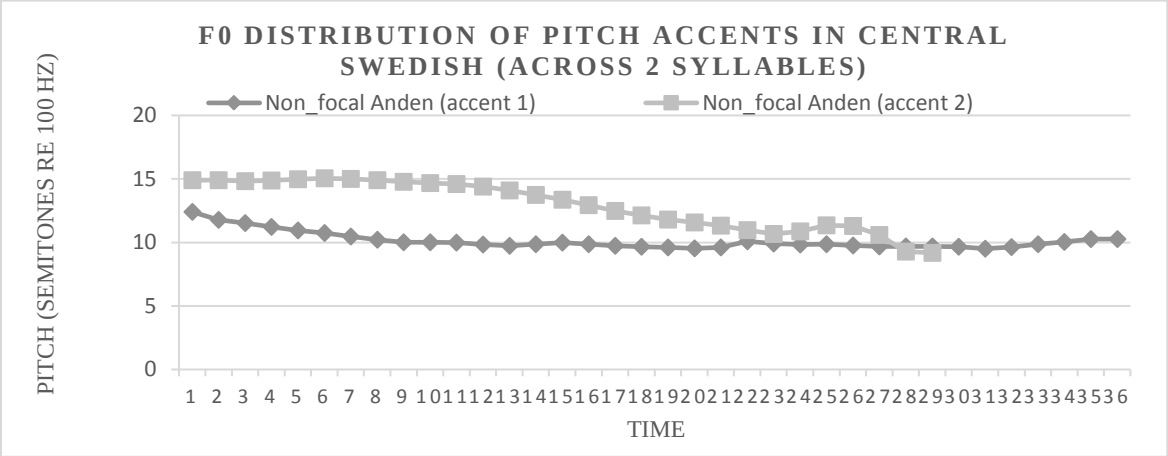


Figure 2.4 F0 patterns (non-focal) in Central Swedish on the disyllabic word *anden* by a female native speaker of Central Swedish in similar age as the native speaker of MC in Figure 2.1. The pitch points were tracked every 10 milliseconds. The duration of *anden* 'the spirit' (accent 2) is slightly longer than *anden* 'the duck' (accent 1), Praat failed to trace all pitch points during the time course of *anden* 'the spirit' (accent 2).

2.2.2.3 Focal accent 1 and 2 in Central Swedish

When accent 1 and accent 2 words in Central Swedish are in focal position in a sentence or phrase, a rise to a high tone occurs after the tonal gesture of the accented syllable. In other words, at the end of the vowel of the stressed syllable, a high tone rises at the vowel of the syllable following after the first stressed syllable in focused accent 2 words. Thus, a focused accent 2 word in Central Swedish has two peaks while a focused accent 1 word in Central Swedish has only one peak (Bruce, 1977) as illustrated in figure 2.5.

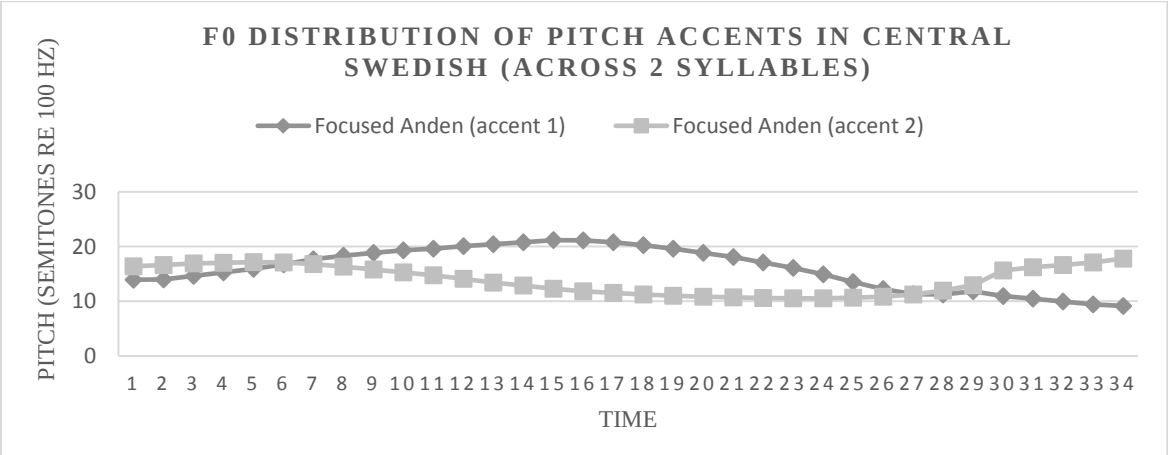


Figure 2.5 F0 patterns (focal) in Central Swedish on the disyllabic word *anden* by a female native speaker of Central Swedish in similar age as the native speaker of MC in Figure 2.1. The pitch points were tracked every 10 milliseconds. A*nden* 'the duck' (accent 1) and *anden* 'the spirit' (accent 2).

2.2.2.4 Online processing of pitch accents in Swedish

Swedish nouns are morphologically inflected for both number and definiteness according to five different declension classes. Accent 1 is associated with single, definite suffix while accent 2 is associated with both plural definite and plural indefinite suffixes and is also assigned post-lexically in compound words. In Swedish verbs, accent 1 is associated with the simple present tense of the verb stem while accent 2 is associated with both the past tense and the simple and past perfect tense of the same verb stem according to four different conjugation classes. Thus, accent 2 in Swedish generally has more competitive suffixes than accent 1.

In their ERPs study of word accents and morphology in Swedish nouns, Roll et al. (2010) found that both correct and incorrect accent 2 suffixes combined with a mismatched accent 1 stem tone produced P600 effects (P600 reflects syntactic violations in stimuli, an indication of reanalysis of incorrect forms), but no increase in the N400 (N400 is typically seen to increase in response to violations of semantic expectations). On the other hand, accent 1 suffixes did not yield any effects in words realized with a mismatched accent 2 stem tone. The results indicate an association between accent 2 tone and its suffixes, and thus further support the assumption that only suffixes co-occurring with accent 2 are associated with pitch accent in mental lexicon while accent 1 is supposed to be assigned postlexically and thus is regarded as a default accent (Riad, 1998, 2009). Furthermore, the study also found that the high stem tones of accent 2 words produced a P200 effect that reflects pre-attentive processing of the potential incoming suffixes and thus increased processing load. In their response time study of processing morphologically conditioned word accents in Swedish verbs, Söderström et al. (2012) found that response time for non-focal accent 2 words with correctly matched accent 2 stem tone were longer than non-focal accent 1 words with correctly matched accent 1 stem

tone. In terms of online processing of focal accent 1 or accent 2, Felder, Jönsson-Steiner, Eulitz, & Lahiri (2009) found firstly, that accent 1 words were recognized more accurately than accent 2 words, and secondly, that response times were longer for accent 2 words compared to accent 1 words when participants were asked to judge between an accent 1 and an accent 2 word after only hearing the focal stem tone (either accent 1 or accent 2) in the initial stressed syllable.

Hence, it seems that accent 1 and accent 2 are processed differently online and the processing load for accent 2 words were longer than accent 1 words in Swedish.

2.2.3 English

English is categorized as an intonation language or as a non-lexical-tone and non-pitch-accent language or as a nontonal language. Pitch variations in English are used to differentiate different sentence types (questions or statements) or to emphasize the new or unpredictable information in a sentence. The use of pitch variations to signal different meanings at the word level in English is thus very limited.

Stress in English is generally accompanied with an increased duration, amplitude, pitch height and vowel quality in the stressed syllable in question. However, changes in F0 were still assumed to be the primary cue in detecting stress in English among the above-mentioned elements in Lehiste (1970). Thus, syllables with high pitch were likely to be perceived as stressed. For example, native English speakers tended to perceive high pitch onset on T1 and T4 in MC as stressed syllable and thus were confused with these two tones in MC (Lehiste, 1970; White, 1981; Shen, 1989).

Take homophonous word pairs such as *SUBject* (noun) and *subJECT* (verb) as an example. In those words, pitch is used together with intensity and duration to indicate different lexical stress at different syllables of word pairs in the way that high pitch is highly associated with high intensity (depending on amplitude) and longer vowel duration of the stressed syllable.

Apart from lower pitch and amplitude and shorter duration of the vowel, the absence of stress in a syllable is also associated with centralization of the vowel of the unstressed syllable (from [ʌ] to [ə] in the case of *SUBject* [`sʌbdʒɪkt] to *subJECT* [səb`dʒɛkt]). Thus, the stressed syllable is perceptually more prominent than the neighbouring unstressed syllable. Figure 2.6 illustrates the waveforms, pitch patterns and intensity patterns for the noun and verb forms of the word *subject* spoken by a female native speaker of English. The speaker exhibited some degree of reductions of vowel duration, lower pitches, and lower amplitudes for unstressed syllables.

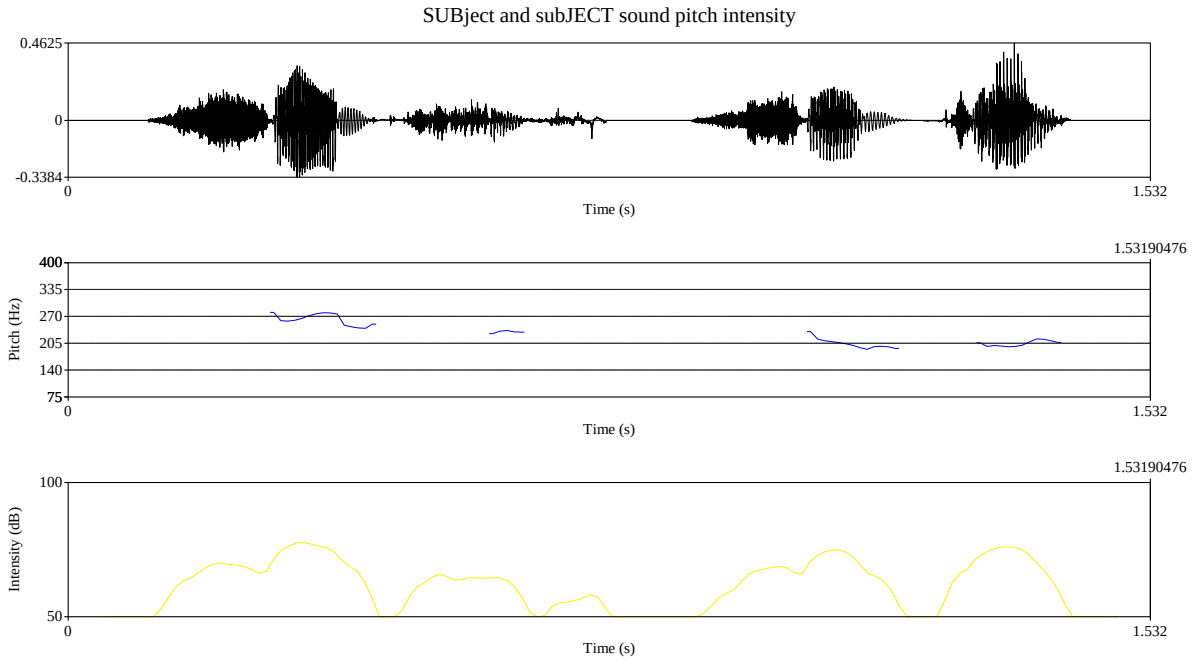


Figure 2.6 Stress patterns on the word *SUBject* and *subJECT* in English by a female native English speaker.

2.2.4 Comparisons among MC, Swedish and English

The stress difference in Swedish can be described in much the same way as the stress difference in English (Ladefoged et al., 2010:260), but stress difference in Swedish does not differentiate word classes in homophonous word pairs (e.g., *CONtract*/*conTRACT* – noun/verb) as often as stress difference in English does[10]. In Swedish the word accent

[10] There are few, countable homophonous word pairs in which stress difference changes word classes in Swedish as well, but not systematically (e.g. beté [béte:] *behave* vs. bete [bè:tə] *bate*).

contrasts are associated with the pitch height of the stem tone of the stressed syllable and there is no tone contrast if the last syllable is stressed. In MC there are also intrinsic stress patterns in disyllabic (primary stress vs. secondary stress pattern) and trisyllabic words (secondary stress-primary stress-unstressed vs. secondary stress-unstressed-primary stress), but an unstressed syllable (stress-unstressed pattern) is always toneless (Guo et al., 2012).

Both non-focal accent 1 and accent 2 have a falling contour in Swedish with non-focal accent 1 beginning with a low pitch height and non-focal accent 2 beginning with a high pitch height in the stem tone of the stressed syllable. In MC both the first half of T3 and the whole T4 have a falling contour with T3 having a low starting point of the pitch height and T4 having a high starting point of pitch height. A pilot study was carried out to investigate categorical perception of two pitch falls (T3 and T4) in MC in which Mandarin-naïve NS listeners were employed as one language group. In the pilot study the pitch points of two pitch falls in MC and Swedish spoken by two female speakers (one native speaker of MC and one native speaker of Central Swedish) in similar age were extracted and plotted in figure 2.7. However, seen from figure 2.7, there are three major differences between the two falling contours in Swedish and MC.

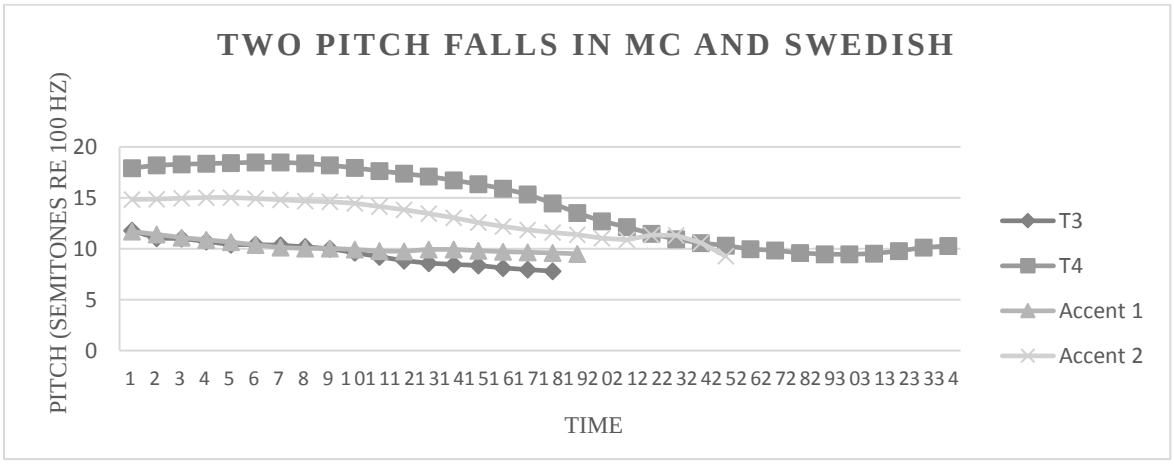


Figure 2.7 The F0 contour of T3 (falling part) and T4 on the syllable *pa* in MC, [ˈa] part of *anden* 'the duck' (accent 1) and [ˈandɛ] part of *anden* 'the spirit' (accent 2) in Central Swedish by female native speaker of MC and Central Swedish in similar age.

First, in MC the pitch range of the pitch fall in T4 is much larger than in T3 whereas the pitch range of the pitch falls of non-focal accent 1 and non-focal accent 2 in Swedish is comparatively similar. Second, with the consideration of T3 having a longer duration than T4, T4 has thus a much steeper descending fall (ΔF/duration)[11] than T3 in MC. In Swedish the fall of non-focal accent 1 is just slightly steeper than the fall of non-focal accent 2 with the consideration that the fall of non-focal accent 1 starts earlier than non-focal accent 2. The duration, pitch range and ΔF/duration of the four pitch falls (two in MC and the other two in Swedish) based on the individual references are listed in table 2.3.

Table 2.2 The duration, pitch range and ΔF/duration of the four pitch falls.

Pitch falls	*Duration (second)*	*Maximum pitch (semitones re 100 Hz)*	*Minimum pitch (semitones re 100 Hz)*	*ΔF/duration (semitones/s)*
T4 (pa)	0.2730	16.64	8.13	31.17
T3 (pa)	0.1891	13.08	10.92	11.42
Accent 1 (a part)	0.0838	11.61	10.16	17.30
Accent 2 (anda part)	0.3036	15.04	11.05	13.14

The third difference lies in that the two falling contours in MC run over a time course of one syllable while the two falling contours in Swedish run over a time course of two syllables. Söderström et al. (2012) found accent 2-associated suffixes yielded generally greater response times than accent 1-associated suffixes for native Swedish speakers (only non-focal word accents were used on test stimuli in the study). The pilot study found that NS listeners demonstrated the highest level of sensitivity to falling pitches when the initial pitch height was at about accent 1 level and the least level of sensitivity to falling pitches when the initial pitch height was high due to lack of corresponding high initial pitch height of the two Swedish accents. However, based on Söderström et al. (2012) and the pilot study, it is

[11] In a narrow sense, "ΔF" here refers to the range of the falling pitch in question (between the maximum pitch to the minimum pitch). "ΔF/duration" here refers to the descending of pitch fall as a function of time (the velocity of the pitch fall).

reasonable to presume, if Mandarin naïve NS listeners perceive disyllabic tokens beginning with T4 (T4T*, *=1 – 4) in MC as non-focal accent 2 words in Swedish, the response time for NS listeners would be longer in disyllabic tokens beginning with T4 (T4T*, *=1 – 4) than disyllabic tokens beginning with any other tones in the present study because of the transfer of L1 background in prosody. And this prolonged response time is assumed not to occur in Mandarin naïve NE listeners because such influence of L1 background in prosody is missing for English speakers.

2.3 Top-down and bottom-up processing in speech word recognition

Speech word recognition in general can be interpreted in two perspectives: bottom-up (also termed as data-based) processing and top-down (also termed as knowledge-based) processing. According to the Trace model (McClelland and Elman, 1986), both types of processing are necessary and cooperate in speech word recognition. Bottom-up processing is a direct acoustic processing of the incoming signal and it happens before the top-down processing which is based on prior linguistic knowledge. The Cohort model (Marslen-Wilson,1987) claims that in order to decode a spoken message, a listener has to extract both segmental and prosodic information from the acoustic and phonetic signal and maps onto presentations of lexical meaning and this speech-meaning mapping involves continuous and dynamic process of activation and competition among multiple, potential candidates.

When a native speaker of MC hears the token *ma* spoken in tone 1 before the incoming context, s/he needs first to recognize the pitch pattern categorized as tone 1 and exclude all the other 16 candidates of *ma* spoken in the other three tones (two in tone 2, six in tone 3, four in tone 4, and 3 in toneless), then with the aid of the incoming context information to exclude the other four candidates of the same segment and the same tone (tone 1), and finally s/he maps the sound with both the segment and the tone representation in the mental lexicon and recognize that *ma* in tone 1 in that situation means *mother*. In short, s/he associates a

particular pitch (or just a cue or several cues of that pitch) associated with a segment first with a certain tone and then with a lexical meaning. Thus, two processes are involved in spoken word recognition in MC: the process of activation of multiple word candidates based on initial speech input (the cohort), and the process of selecting the one among the competitors that maps best with the lexical meaning in mental lexicon. The process of selection is probably primarily driven by bottom-up phonetic and acoustic inputs and then adjusted by top-down lexical representation.

However, the token *ma* has just few homophonic candidates and its lexical processing time is relatively short. There are other tokens in MC that have apparently much more homophonic candidates. For instance, the token *yi* has 148 lexical tone candidates (80 in tone 1, 20 in tone 2, 28 in tone 3 and 20 in tone 4) and the time to extract all competitive homophonic candidates for token *yi* is huge if the process of lexical tone works solely in such a temporal way. Thus, syllables with many competitors are dependent on larger cohorts (more than one syllable) to disambiguate from other homophonic candidates and to pre-activate the incoming speech signals in order to map the lexical presentation in mental lexicon. During this rapid process of spoken word recognition, the activation of candidates with low frequencies will be overridden by the incoming speech signals due to their longer activation time.

Likewise, pitch accents in Swedish are determined by the morphological characteristics of the word in form of different suffixes. All suffixes induce a stem tone onto the word. What differs in the function of the tone between Swedish and MC is that the tone in Swedish (accent 1) itself is not directly associated with the lexical meaning of the word, but in which suffix it is attached to. Thus, prosodic information in pitch accent language such as Swedish is rather to pre-active suffix (Roll, Horne & Lindgren, 2011; Roll et al. 2015, 2017).

Native speakers of non-lexical-tone and non-pitch-accent language such as English in which prosodic information plays a minor role in word recognition compared to segmental

competitors, probably rely on multiple cues such as pitch, intensity and duration to differentiate minimal pairs in stress pattern such as *INsert* and *inSERT*.

Nevertheless, humans often predict incoming signals based on experience. The top-down processing of pitch height of the stem tone in predicting the coming suffix of the word in Swedish can possibly be transferred or extended in perceiving the non-native tones in another character and scope due to linguistic experience (Schremm et al., 2016). On the other hand, native speakers of a non-lexical-tone and non-pitch-accent language such as English are simply unable to relate lexical tones to familiar native representations of such a prosodic category because there is nothing in their native "prosodic grammar" that prepares them to decode a prosodic category such as changes in pitch at the word level in a morphologically (or lexically) contrastive manner. Thus, this phonological awareness is missing in native speakers of English. However, as mentioned earlier (Chardrasekaran et al., 2010; Wong et al., 2007) this phonological awareness can be trained through pitch-to-meaning training paradigm for native English speakers and thus contribute to learning of a tonal language. For native speakers of Swedish, it is rather a question of training the extension of scope and scale of phonological awareness of the top-down processing of pitch height they possess.

2.4 General influencing factors in pitch perception

2.4.1 Psychoacoustic ground for pitch perception

From a psychoacoustic perspective and based on his frequency discrimination study, Small (Small, 1973:375) found that differential perceptual threshold for frequency was influenced by stimulus duration. The differential perceptual threshold showed a nonlinear contra-variant relationship with stimulus duration if the duration was shorter than 200ms as illustrated in figure 2.8.

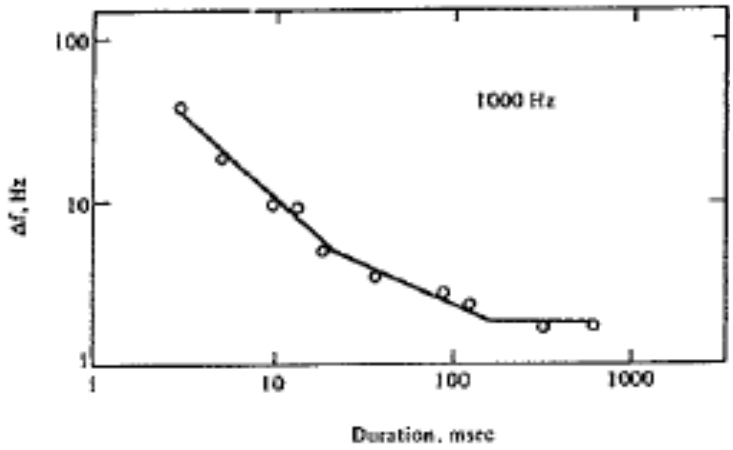


Figure 2.8 Differential threshold for frequency as influenced by stimulus duration (Small, 1973:375).

From the same perspective, Klatt (1973:8) found that the minimal detectable difference in pitch was in the region of 0.3 Hz in a constant F0 of 120 Hz (a level tone), but the minimal detectable difference in pitch increased to 2 Hz instead in a F0 with a 120 Hz as starting point but with a linear descending ramp (32Hz /sec) (a falling tone). It was unclear whether the enlarged discrimination threshold was caused by the rate of pitch falling or the falling pitch per se, however, it was concluded from the study that it was more difficult to detect an average F0 change if the change occurred in a descending ramp F0 contour.

Zee & Greenberg (1977) found that if the duration of a syllable is less than 40-65 ms, the pitch contours were hard to perceive. As mentioned above, the duration of T4 (270 ms in syllable *ma* in the practice session of the experiment and 273 ms in syllable *pa* in the real sessions of the experiment) was shortest among all other tones in MC. Bearing in mind that both shorter duration and falling pitch (in contrast to level pitch) enlarges the JND (just-noticeable difference) in pitch perception from a psychoacoustic perspective, it is rational to anticipate that T4 might cause perceptual confusions for both NS and NE listeners.

2.4.2 Music-to-language transfer

As illustrated in figure 2.9, the hearing range of speech sound is marked in a form of "banana" in the centre of human audibility and the hearing range of music is marked in a form of "ellipse" including the "speech banana" and the largest circle is the range of human audibility. The useful range of human audibility is usually to be the area between the individual's absolute threshold of hearing (audible) and the threshold of feeling (to reach auditory pain threshold) (Denes & Pinson, 1993:96).

The hearing range of music (music domain) is broader in terms of both pitch and intensity range compared to the hearing range of speech (speech domain). Thus, the feeling to hear human singing is generally more dynamic than to hear a human speaking (Lindblad, 2000:8). As mentioned in 1.4.2, the pitch range in MC is reported as wider than a series of languages such as Japanese, Cantonese, Thai, English and Swedish, thus MC sounds consistently more dynamic in the sense of pitch range than the above-mentioned languages. Furthermore, MC sounds more dynamic in terms of intensity range compared to Swedish and English as well in the sense that both creaky phonation (low intensity) and modal phonation are included in MC and the quick energy drop owing to the steep slope of pitch of T4 in MC. Therefore, it is reasonable to presume that participants with higher musical aptitude will have a higher performance in perceiving tones in MC due to the level of dynamics in terms of both pitch and intensity range in MC has a relatively larger overlap with the hearing range of music compared to the other aforementioned languages.

In addition, both behavioural (Wong et al., 2007) and neurophysiological studies (Bidelman et al., 2010, 2011) found that musical training could enhance language-related pitch processing (music-to-language transfer), but converse results (language-to-music transfer) were not found – neither native speakers of a tonal language (MC) nor non-musicians were able to discriminate subtle changes in musical pitch with the same accuracy as musicians (Bidelman,

et al., 2011). However, Shen (2016) found that for native speakers of a tonal language (MC) who had already developed a mental template in which a lexical meaning was associated with a certain tone, musical training did not have a significant impact on identification of native tones as well as pitch-transposed tones when pitch height was manipulated. Hence, the present study provides a unique opportunity for investigating how musicality dynamically influences non-native perception of tones in language due to the fact that participants were native speakers of a non-lexical tone language and their musical aptitude was reported in a five-degree scale.

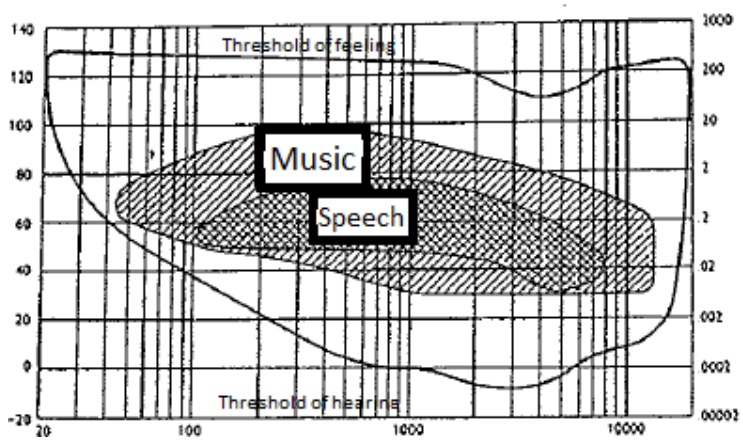


Figure 2.9 Human audibility – pitch and energy distribution in speech and music (Hadding & Petersson, 1970).

2.5 Perceptual learning of lexical tones

2.5.1 The effects of perceptual training

Aside from the influence of L1 background in prosody that has been introduced in the introduction part of this thesis, previous studies have also reported the effectiveness of short-term laboratory perceptual training on the learning of non-native sound (Leather, 1990; Wang, Spence, Jongman, & Sereno, 1999; Wang, Jongman & Sereno, 2003; Wayland et al., 2004, 2008; Francis et al., 2007; Wong, et al., 2007; So, 2006; So et al., 2010; Chandrasekaran et al., 2010) in terms of the ability to identify and discriminate non-native lexical tone contrasts among adult learners. The scale and degree of learning, however, differs from study to study due to differences in characteristics of learners (naïve learners vs. experienced learners of

different proficiency level in the studying language in question), stimuli type (natural vs. synthetic; monosyllabic vs. multisyllabic tokens) and training procedure, among others. The training procedure differs in many aspects in the aforementioned studies, among which two aspects are relevant to this study: traditional auditory training paradigm or sound-to-meaning training paradigm; identification and/or discrimination training procedure.

Using the high-variability auditory training procedure, Wang et al. (1999) auditorily trained native American learners of Mandarin Chinese (with one or two semesters' learning experience of MC) to identify the four tones in MC in two weeks. The results showed that the trainees' identification accuracy increased by 21% averagely from the pretest. The identification accuracy did not improve for the control group without auditory training.

Wayland et al. (2008) employed two perceptual auditory training (either an identification [ID] or a categorical same/different discrimination [SD]) procedures to increase Thai-naïve native English and native Chinese speakers' ability to discriminate the mid- vs. low-tone contrast in Thai under two ISI conditions (500 and 1500 ms). The results showed that the mean percentage gain was between 8.8% and 4.2% in ID training and between 7.5% and 5.2% in SD training for native Chinese speakers, and the mean percentage gain was between 14.4% and 12.9% in ID training and between 7.7% and 9.4% in SD training for native English speakers. The numerically higher percentage gain for ID training compared to SD training was explained that focusing on the inclusionary or grouping strategy of ID training facilitated categorical learning. The reduced increase in percentage gain in identification training for the native speakers of English compared to Wang's (1999) result (21%) was explained as less training time.

Francis et al. (2007) employed an auditory training paradigm with closed response-set identification and pairwise difference rating tasks as measurement to investigate perceptual learning of Cantonese lexical tones by Cantonese-naïve native Mandarin and native English

speakers. The results showed that the identification accuracy increased with 9.3 percent units for native speakers of Mandarin and with 16.7 percent units for native speakers of English. There was no upper time limit on participants' response and response times were not recorded in this experiment.

Wayland et al. (2004) investigated the effects of L1 background and different models of ISI (500 ms vs.1500 ms) on perception (both identification and discrimination) of a mid-versus a low-tone contrast in Thai by Thai-naïve native English and native Chinese speakers both before and after auditory training. The results showed that the performance in identification was generally better than the performance in discrimination for both language groups and both ISI conditions.

In Wang et al. (1999), Wayland et al. (2004; 2008) and Francis et al. (2008), the participants were asked to focus their attention on the pitch level or the tone of the word. Wong et al. (2007) focused on training participants to use pitch patterns for lexical identification and thus increase phonological awareness of pitch patterns at word level for speakers of a non-tonal language. By employing a phonetic–phonological–lexical continuity training paradigm, the authors investigated Mandarin-naïve native American speakers' ability to learn to use pitch patterns to identify a vocabulary of six English pseudosyllables superimposed with three pitch patterns (totally 18 words) and found that successful learners took 7.22 sessions (30 minutes per session) to reach identification accuracy of 95% correct level or 5% improvement (between sessions) for four consecutive sessions and less successful learners took 9.38 sessions to achieve the same identification level or the improvement between sessions. The learners' identification performance in this study was not time-based just as Francis et al. (2007). It was worth noting that the participants' correct nonlexical pitch identification before training was 78.05%. The authors thus raised the issue of the importance of a phonetic–

phonological–lexical continuity including phonological awareness in adult nonnative tone perceptual learning.

2.5.2 Issue of ID and SD training procedure in tone perceptual training

In Wang et al. (1999) and Wong et al. (2007) the ID training procedure was adopted, in Wayland et al (2008) both ID and SD training procedure were employed and in Wayland et al (2004) ID and SD training procedure were compared as a training paradigm. In Kaan et al. (2008) and Lu, Wayland & Kaan (2015), only a behavioural SD training was suitable based on the characteristics of mismatch negativity (MMN) studies. In perceptual study of non-native sounds, the issue of the effectiveness of ID and SD training procedure has been a subject of debate.

Concerning human discrimination of sounds, Lindblad (2000) describes: "In the process of sound perception, we identify and categorize different sounds when we first hear them rather than discriminate the difference of the two sounds" (p. 10). Jamiesson & Morossan (1986) argued that SD training procedure encouraged participants to pay more attention to acoustic between-category differences than to acoustic within-category differences and thus the participants failed to recognize the stimuli with slightly acoustic differences as the same category. Lively et al. (1994) focused the adaption of immediately feedback after every ID training trail in promoting the match between the new learned acoustic categories and the acoustic codes stored in long-term memory.

Based on the perceptual learning studies on non-native sounds and the arguments of between ID and SD training procedure mentioned above, the present study will focus on four aspects in perceptual learning of tones in MC. Firstly, a phonetic–phonological–lexical continuity training paradigm will be adopted with the purpose of raising participants' phonological awareness of lexical tones. Secondly, ID training procedure will be employed. Due to this study investigates online learnability of lexical tones at a naïve level, 100% ID training

procedure would not be applicable. Thus, a mixture of 75% matched trials (sounds match images) and 25% mismatched trials (sounds mismatch images) in behavioural tasks will be employed[12]. Thirdly, a simple speaker instead of multiple speakers in sound stimuli and a simple immediate feedback will be given after every response to promote the match between the new learned acoustic category and the corresponding lexical meaning in the learning language and to avoid participants being adversely affected by acoustic variations (in terms of different speakers) that are irrelevant to the identification of new acoustic categories. Fourthly, the ITI (inter-trial-interval) in this study will be longest as 4500 ms (depending on participants' reaction time) and participant's reaction time was measured and used as a secondly criteria of participants' learning performance. Furthermore, Leathers' (1990) criteria of 80-85% correct identification level for successful learning will be used as the judgement for learnability for the present study.

2.6 The present study

As mentioned earlier in the introduction, this study was conducted to investigate the influence of a pitch accent prosodic system (Swedish) on perceptual learning of a prosodic system of a lexical tone language (MC) at naïve level and to evaluate the relative efficacy of an auditory-image (AI) training approach in a short-time laboratory training. Two research questions and their relevant hypotheses will be tested and explained more extensively in 2.6.1 and 2.6.2. Other relevant elements that might generally influence the perceptual learning are also raised briefly in this section.

[12]Excluding of mismatched trials would not be applicable due to the nature of online-learning procedure in this study (instead of pretest-training-posttest procedure as in aforementioned previous studies). If only matched trials (including only identification) are adopted in this study, the only choice for participants is to "agree" that the sound stimulus matches the image stimulus in every trial. The unvaried training procedure has the potential to lead to the equally optimal outcome of the training (100% correct identification) for both language groups without active learning. 75% matched trials in the behavioural tasks will lead participants' main attention to categorical learning, and minor attention to between-category acoustic differences in stimuli in order to decrease potential monotonousness of the training procedure in question. More details of the design list of stimuli are described in the method part.

2.6.1 RQ 1: The influence of L1 prosodic background

The first research question is: How does language background affect naïve and non-native perception of lexical tones?

Based on earlier cross-linguistic studies on lexical tone perception reviewed earlier in this study (Burnham et al., 2015; Schaefer et al., 2014; So, 2010), it is hypothesized that native listeners of Swedish (a pitch accent language) have an advantage over native listeners of English (a nontonal language) on the perception of tones in MC (a lexical tone language) with which they have no prior experience. Due to experience in tonality use at word level and the tone-meaning awareness are missing for NE listeners, it is expected that NS listeners will outperform NE listeners with a higher general credits of response accuracy in the behavioural task in this study.

It is also hypothesized that NS listeners could be both benefited and restricted in naïve perception of tones in MC with the considerations of the similarities and differences between the two languages (So, 2010; Dong et al., 2013). Flege's (1995) SLM (speech learning model) considers that a native category that is similar but not entirely identical to a non-native one might actually interfere with the perception of a non-native tone (Francis 2008, p.273). Thus SLM predicts increased difficulty for NS listeners in perceiving tone combinations with T4 as initial tone [T4T*(* = 1-4)] in MC if NS listeners perceive T4T*(* = 1-4) in MC as similar to accent 2 words in Swedish and trigger an accent 2 perception and thus relatively lower response accuracy and longer response time compared to tone combinations with T3 as initial tone [T3T*(* = 1-4)] (Söderström et al., 2012; Felder et al., 2009).

It is also expected that both NS and NE listeners might use other perceptual cues besides F0 in perceiving tones in MC. For example, the amplitude contours for T1 and T2 were similar (though pitch contours were different for T1 and T2) in the experiment and the stress differences in Swedish and English are similar as well, thus it is reasonable to presume that

both T1T2 and T2T1 will be difficult for both language groups. Furthermore, T3 differs in both duration and phonation mode compared to the other three tones in MC. Hence tone combinations [T3T* (* = 1-4)] are presumed to be easier to perceive compared to other tone combinations if participants use duration and phonation mode either as primary or as secondary perceptual cue, or as both. In addition, NS listeners are expected to have higher response accuracy and lower response time in perceiving tone combinations with T3 as initial tone T3T*(* = 1-4) in MC than T4T*(* = 1-4) if the low initial tone in T3T*(* = 1-4) triggers an accent 1 perception (Söderström et al., 2012; Felder et al., 2009).

2.6.2 RQ 2: The efficacy of AI training paradigm

The second research question is: Is auditory-image (AI) training paradigm effective in naïve listeners' perceptual learning of tones in MC?

It is hypothesized that a phonetic–phonological–lexical continuity training paradigm will be more effective in extending or raising participants' phonological awareness of lexical tones and thus facilitate perceptual leaning of lexical tones in terms of percentage of correct response compared to traditional auditory training paradigm reviewed earlier in this study. It is expected that the performance in mean response accuracy will be improved across blocks for both language groups and the NS listeners will get greater improvements at the very early block than the NE listeners due to their intrinsic possession of tone-meaning awareness to some degree. It is also expected that the performance in mean response time will decrease across blocks for both language groups as further evidence of perceptual learning. It is also expected that RA will be lower and RT will be longer in mismatched trials than matched trials for both language groups according to Lindblad's (2000) presumption on perception of new sounds.

2.6.3 Other factors of note

Interaction between effects of L1 background and AI training paradigm, the influence of musical aptitude on perceptual learning and psychoacoustic ground for pitch perception will be mentioned shortly in this section.

Both research questions and hypotheses in this study focus on the effects of two factors – L1 prosodic background and AI training paradigm on perceptual learning of lexical tones in MC. However, it is possible that one language group can be more benefited from the AI training paradigm than the other language group. It is expected that the language group hypothesized to have lower learning performance (NE listeners) will benefit more from the AI training paradigm because NE listeners have no experience of using pitch variations to differentiate meanings at word level. If the opposite results are found, it is presumed that AI training paradigm is not effective in raising tone-meaning awareness for NE listeners, but rather more effective in extending tone-meaning awareness for NS listeners with the consideration of the number of matches between tone combinations and images (16).

As for the influence of musical aptitude on perceptual learning, it is expected that higher musical aptitude will give participants a better start baseline and will speed up the perceptual learning across training blocks.

From the perspective of psychoacoustic ground for pitch perception in general, it is expected that T4T*(* = 1-4) will cause perceptual confusions for both language groups due to its comparably short duration and steep pitch fall.

3. Method

This part describes the experimental methodology for the present study. It provides information about the participants, the sound and image stimuli, the design list of the stimuli, the experimental design, the procedure of the experiment and the data analysis.

3.1 Participants

Ninety-two adults participated voluntarily in the present experiment and received a movie ticket for their participation. They were recruited based on three criteria: They were either native speakers of Swedish or English, they had neither earlier experience of studying any tonal languages nor been exposed to any tonal languages earlier (naïve listeners of MC) and the age range for participation was between 18 and 40. The participants were divided into two groups based on their native languages: English and Swedish. The Swedish participants (N = 33, 15 Male and 18 Female) ranged in age between 20 and 34 years (M = 24.81, SD = 3.81) and all were born and raised in Sweden. The English participants (N = 33, 15 Male and 18 Female) ranged in age between 19 and 40 years (M = 25.18, SD = 5.41) were born and raised in their home countries (The United States, United Kingdom, Canada and Australia). All Swedish participants were students at Lund University. Twenty-eight English participants were students (master students or exchange students) at Lund University and 5 English participants worked as English teachers at an international school in Lund. No participants had any form of reading difficulty that could cause them problem in reading instructions of the experiment. Previous experience of musical training was not excluded as a criterium in recruiting participants due to the difficulty in recruiting participants of native English speakers in Lund. All participants were asked to report the degree of their musical aptitude – a factor known to enhance lexical tone perception (Wong et al., 2007) – in a five-degree form in a background information paper after the experiment (see Appendix 1). The scales of musical aptitude of the two language groups were taken into consideration in data analysis. All

participants were informed about principles of research ethics guidelines by the Swedish Research Council before making decision on participation. The first 20 participants were excluded from data analysis due to technical problems. One participant has no hearing on the right ear and was excluded from the data analysis. All the other participants had normal vision and hearing according to self-report.

3.2 Stimuli and design

3.2.1 Sound stimuli

Two monosyllabic tokens *pa* [p^{h}a] and *ta* [t^{h}a] were used as basic syllable structures of the sound stimuli in the experiment. The vowel *a* [a] was used as the vowel of the syllable because of its high ranking in terms of sonority hierarchy. In addition, the pitch change is easier to perceive if it is imposed on a steady-state vowel such as a monophthong than on a non-steady vowel such as a diphthong or a triphthong, in which formants change during the vowel transition of the sonorant part of the syllable (Klatt, 1973). The choice of voiceless stops *p* and *t* as consonants of the syllable was to make the concatenating of the syllables easier. Both *pa* [p^{h}a] and *ta* [t^{h}a] are legal and simple syllable structures in all three languages, thus the participants only need to focus their attention on the lexical tones per se without being distracted by unfamiliar syllables during the experiment process.

Monosyllabic tokens *pa* [p^{h}a], *ta* [t^{h}a], and *ma* [ma] with 4 tone types respectively were pronounced twice in citation form by a 20-year-old female native speaker of MC in the anechoic chamber at the Humanities Laboratory, Lund University. The tokens were recorded at 44100 sampling rate with a 16-24 bit amplitude resolution. The two monosyllabic tokens *pa* [p^{h}a] and *ta* [t^{h}a] with 4 tone types respectively were later systematically concatenated into 16 disyllabic tokens with 16 possible tone combinations (using Audacity version 2.1.3) and used as sound stimuli in the experiment. The 16 tone combinations of disyllabic token *pa ta* [p^{h}at^{h}a] were also recorded by the same speaker and used as references for the concatenation of the monosyllabic token *pa* [p^{h}a] and *ta* [t^{h}a]. The sound stimuli were judged as highly

natural by two native speakers of MC except for the stimuli with T3 as the first syllable because in natural speech T3T*(* = 1-4) go through T3 sandhi.

The durations of the monosyllabic token *pa* [pʰa] with four tone types respectively were listed as follows: 429 ms for *pa1* [pʰa^{55}], 462 ms for *pa2* [pʰa^{35}], 489 ms for *pa3* [pʰa^{214}] and 426 ms for *pa4* [pʰa^{51}]. The durations of the monosyllabic token *ta* [tʰa] with four tone types were 391 ms for *ta1*[tʰa^{55}], 412 ms for *ta2* [tʰa^{35}], 471 ms for *ta3* [tʰa^{214}] and 359 ms for *ta4* [tʰa^{51}] respectively. The two monosyllabic tokens *pa* [pʰa] and *ta* [tʰa] were aligned with each other in such a way that the time window of the first monosyllabic token [pʰa] was kept at 600 ms regardless of its original durations with different tone types respectively[13] and the second monosyllabic token [tʰa] kept its original duration with the four tone types. The time window of the first syllable of the disyllabic token was kept at 600 ms regardless of its original duration in different tone types in order to make the counting of response time (RT) as a cue of tone discriminations easier. For example, if RT is 648 ms for one trial, it can be supposed that the participant makes her/his judgement based on the pitch onset, rather than the whole pitch contour, of the second syllable of the disyllabic token in question. The actual duration was 991 ms for *pa* (*=1-4) ta1*, 1012 ms for *pa* (*=1-4) ta2*, 1071 ms for *pa* (*=1-4) ta3* and 959 ms for *pa* (*=1-4) ta4* respectively. Figure 3.1 shows the waveforms, the pitches and the durations of the four tone combinations with *pa4* [pʰa^{51}] as the first syllable.

[13] A silence of 171, 138, 111 and 174 ms was inserted between pa1, pa2, pa3, pa4 and ta1-4 respectively.

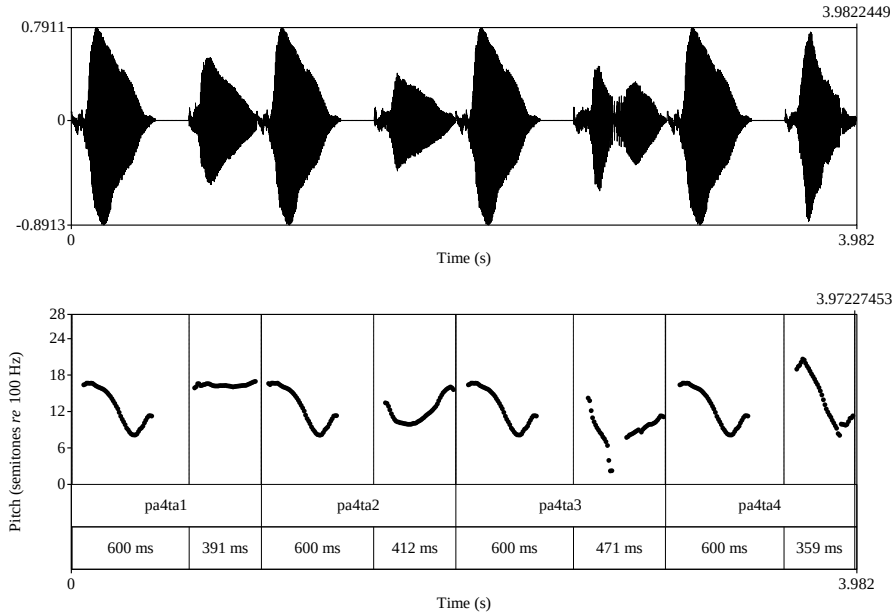


Figure 3.1 The waveforms and pitches of the four tone combinations with *pa4* [p^ha^{51}] as the first syllable (a silence of 174 ms was inserted between pa4 and ta1-4).

3.2.2 Image stimuli

For every disyllabic tone combination (token) there was a corresponding lexical meaning in form of an animal image. Altogether 16 different disyllabic tone combinations matched with 16 different animal images and employed as the sound stimuli and the image stimuli respectively for the experiment. The underlying idea of the corresponding matches between the tone combination and the animal image was that the lexical meaning changes when the tone of one syllable of a disyllabic token changes in a lexical tone language. For example, the sound stimulus pa1ta1 (T1T1, tone 1 + tone 1) was matched with a duck and the tone combination pa1ta2 (T1T2, tone 1 + tone 2) was matched with a chicken. These two sound stimuli had the same syllable structure and differed only in the onset of the tone of the second syllable. In a lexical tone language this tiny difference in tones changes the meanings of two words otherwise exact the same.

When a participant heard a tone combination pa1ta1 and saw a duck on the computer screen, s/he had to guess whether pa1ta1 in MC meant duck by pressing a special key (1-key = agree and 2-key = disagree) at the first time and s/he was informed about the guessing result

immediately through a simple feedback. Next time when the tone combination pa1ta1 was heard and the duck appeared on the screen again, the participant would press the 1- key if s/he had learned pa1ta1 in MC means duck (a matched trial). When the tone combination pa1ta1 was matched with a chicken, s/he would know that the tone combination pa1ta1 and a chicken was a mismatch and thus press the 2- key (a mismatched trial). The correct matches were categorized as matched trials (only including identification) and the mismatches were categorized as mismatched trials in data analysis. Apart from the function of decreasing potential monotonousness with 100 % ID training procedure, the adoption of the mismatched trials was also to promote participants' auditory performance in differentiating tiny differences between different tone combinations.

3.2.3 Design list of stimuli

There were 6 wrong possible tone combination variations for every correct tone combination. Taking disyllabic tone combination pa1ta1 as an example, 3 wrong tone combinations with the first syllable – the initial tone – variations were pa2ta1, pa3ta1 and pa4ta1 respectively and 3 wrong tone combinations with the second syllable – the final tone – variations were pa1ta2, pa1ta3 and pa1ta4 respectively. The design lists with both the correct and the incorrect matches between the sound stimuli and the image stimuli for the whole experiment are illustrated in Appendix 2.

Every correct-matched tone combination and animal image ran 18 times during the experiment and the total number of trials for correct-matched tone combination and animal image were 288 (16×18). The mismatch variations between tone combination and animal image for every tone combination were 6 and the mismatched trials ran 6 times for every tone combination during the experiment, thus the total number of trials for mismatched tone combination and animal image were 96 (16×6). The percentage of training trials with correct-matched tone combination and animal image during the whole experiment was 75%. The

experiment part ran 384 trials totally in one experiment procedure. The trials were divided into 6 blocks and in each block 64 trials ran. In each block the proportion of matched trials and mismatched trials was 3:1. The presentation of the stimuli was counterbalanced across all training blocks and participants. The distribution of all variations of mismatch between the sound stimuli and the image stimuli was thus not distributed evenly block-by-block in this experiment. For example, the altogether six mismatches between the sound stimuli T1T1 and six different "mismatched" animals might be distributed twice in block 1 and none in block 2 and so on. The experiment was edited using E-Prime.

3.2.4 Experimental design

The design of the experiment was a 16×2×6×2 factor design, that is tone combination (16 levels:T1T1, T1T2, T1T3, T1T4, T2T1, T2T2, T2T3, T2T4, T3T1, T3T2, T3T3, T3T4, T4T1, T4T2, T4T3, T4T4) × trial (2 levels: matched vs. mismatched) × block (6 levels: block 1, block 2, block 3, block 4, block 5, block 6) × L1 background (2 levels: Swedish vs. English). Among these, tone combination, trial and block were within-subjects factors and L1 background was a between-subjects factor.

The independent variable of the experiment was that 16 tone combinations contained 16 different lexical meanings (in form of 16 animal images). Thus, the correct match between the tone combinations and the corresponding animal images was manipulated and defined as matched trials and the mismatch between the tone combinations and the corresponding mismatched animal images was manipulated and defined as mismatched trials.

The dependent variables were response accuracy (RA) and response time (RT) of matched and mismatched trails for the 16 tone combinations.

3.3 Procedure

After signing an informed consent form, a short introduction about how the tonal system works in MC in general and the details of the experiment procedure were given to every

participant orally by the same experimenter before the experiment. The experiment was performed by participants individually in a quiet room in a session with one practice part and a six-block experiment part. Between the practice part and the experiment part and between every block of the experiment part, the participants were given the chance to take a break or to ask questions to the experimenter who was sitting outside the experiment room. The experiment was performed using E-Prime. All participants were coded numerically and their identity could not be traced back.

The participants were seated in front of a computer screen and listened to the sound stimuli while watching the image stimuli. The sound stimuli were presented binaurally through headphones at a constant and comfortable level. Instructions written in English were presented on the computer screen before every procedure the participants were expected to undergo[14]. As explained earlier in section 3.2.2, the participants just guessed if there was a match or mismatch between the sound stimulus and the image stimulus at the first time. They learned if they had guessed right or wrong through feedback and in this way they were expected to learn eventually some of the names of animal images during the training process. After the experiment, the experimenter debriefed the purpose of the study to participants and explained the details of the experiment to those participants who showed interests.

The experiment consisted of two parts. The first part was the exercise part in which only monosyllabic token *ma* with T1 and T3 were presented as sound stimuli and the images of *mother* and *horse* were presented as image stimuli. The percentage of correct match and incorrect match between the sound stimuli and the image stimuli was 50% respectively. The participants were informed that they were encouraged to practice the exercise part as many times as needed until they felt acquainted with the paradigm.

[14] All Swedish participants have a high proficiency of English.

The experiment part consisted of six blocks. In each block 64 trials were run. The duration for image stimuli was set at 2000 ms, the duration for sound stimuli was set at 3000 ms with 300ms pre-release time and the feedback duration was set at 1500 ms. After every response, a simple feedback indicating a correct or an incorrect response was given followed by the next presentation of the stimuli. If it was the incorrect response, the correct answer was not given. The participants could not replay a stimulus. All stimuli scrolled on the computer screen in the time course of one block. The participants were informed to feel free to take a pause between the blocks during the process of the experiment. The ITI was longest as 4500 ms and the whole experiment lasted about 28.8 minutes with the time for the practice part and the pause between the blocks excluded.

3.4 Data analysis

The mean values of response accuracy (RA) and response time (RT) were measured for testing the effects of L1 background and learning performance of both language groups. The RA and RT data were statistically analysed using SPSS 25.0. Chiefly only significant results are reported in the "results" part of this study. Occasionally, non-significant results are mentioned for comparison.

3.4.1 Response accuracy (RA)

The mean value of RA was calculated in terms of the average of the sum of correct response (value = 1) and incorrect response (value = 0).

Total score (mean value of RA) for all trials for each participant in every language group was obtained and submitted to 16 (type of sound stimulus: tone combination) × 2 (L1 background: Swedish vs. English) repeated measures ANOVA without any covariates, with musical aptitude and age as covariate once at a time, and with both musical aptitude and age at the same time as covariate(s) to examine and compare the overall performance of both language groups for the whole experiment and to examine how covariates such as musical aptitude and

age influence the variables by controlling them and then including or excluding them. The alpha level was set to .05. The effect of between-subjects L1 background was to test whether the two language groups differed significantly from each other in results of mean values of RA for all trials in terms of all tone combinations.

The influence of age was not the focus of this study, however, due to the age range of participants and the burden of auditory memory capacity[15] were relatively large in this study, it was thus unavoidable to examine the degree of its influence on learning performance. If the influence was negligibly small, the age as covariate would be excluded as covariate from data analysis. On the other hand, musical aptitude was anticipated to boost lexical tone perception based on earlier similar studies (Burnham & Brooker, 2002; Alexander, Wong & Bradlow, 2005; Wong et al., 2007). Thus, in all treatment conditions in this study, musical aptitude was counted as covariate. It was examined how different variables were influenced by musical aptitude compared to without, but not reported in data analysis due to musical aptitude being treated as a confounding factor instead of a within-subjects factor in this study. Only change or disappearance of effects of variable(s) caused by counting musical aptitude as covariate, compared to without counting it as covariate, was mentioned in the summary of improvements across blocks and effects of L1 background on perceptual learning of tones in MC.

To test effects of sound and image match/mismatch on learning performance, the mean values of RA for each participant in each group were submitted to a 2 (type of trial: matched vs. mismatched) × 16 (type of sound stimulus) × 2 (L1 background) repeated measures ANOVA, with type of trial and tone combination as within-subjects factors, L1 background as between-subjects factor and both with and without musical aptitude as covariate. The effect of within-

[15] Participants must remember the names of 16 animals in form of different tone combinations both quickly and precisely because the ITI was only 4500 ms and replay of the stimuli was not allowed in the experiment. It raises higher demands on verbal memory capacity for relatively less young participants.

subjects factor trial was to test how learning performance differ between matched and mismatched trials, thus further assess the participants' identification ability as naïve listeners as a result of training.

To examine two language groups' block-by-block improvements, the mean values of RA for each participant were submitted to a 6 (block) × 16 (type of sound stimulus) × 2 (L1 background) repeated measures ANOVA, with block and tone combination as within-subjects factors, L1background as between-subjects factor, and both with and without musical aptitude as covariate in all and matched trials. The effects of within-subjects factors were conducted by each language group separately with and without musical aptitude as covariate to examine whether musical aptitude could facilitate perceptual learning. In addition, the mean values of RA were listed for all tone combinations for each language group and the easier tone combinations and the confusing tone combinations for each language group across blocks were compared.

As mentioned earlier in the introduction part (1.4.3), this study was a systemic learning of tones in MC for naïve listeners in which 16 tone combinations were used as sound stimuli. In order to test the effects of different tones depending on their position in the tone sequence and make it easier to compare tones in disyllabic tokens in MC with pitch accents in Swedish as mentioned in the introduction part (1.4.2), an additional analysis was carried out, where the 16 tone combinations were also reorganised into four sub-treatment conditions (four-treatment design) in terms of tone combinations with the same tone at the first syllable and with the same tone at the second syllable distinctly. In this way, a more detailed and convincing demonstration of a cause-and-effect relationship between the independent and dependent variables resulted (Gravetter & Forzano, 2018:273). The sub-treatment condition of tone combinations with the same tone of the first syllable were T1T*, T2T*, T3T*, and T4T* (* = 1-4) and termed as "initial tone sub-condition", and the sub-treatment condition of tone

combinations with the same tone of the second syllable were T*T1, T*T2, T*T3 and T*T4 (* = 1-4) and termed as “final tone sub-condition”. To examine which initial or final tone(s) was/were significantly easy or difficult compared with other initial or final tone(s) in their corresponding sub-treatment condition, estimated mean values of pairwise comparisons for matched and mismatched trials separately were conducted for each language group once at a time. The musical aptitude was referred to as a covariate to examine how musical aptitude influenced effects of tone combinations in both initial and final tone sub-condition.

3.4.2 Response time (RT)

The mean value of response time (RT) was statistically analysed by using SPSS 25.0 in a similar way as the analysis of mean value of response accuracy (RA).

The mean response times were computed for all correct responses for each participant. Incorrect response, no response and response times less than 600 ms were excluded from calculation. Response time less than 600 ms means that the participant had made a judgement before listening to the second tone of the tone combination. The response time was measured from the onset of the first tone of the tone combination in each trial.

The increase in RA and the decrease in RT across blocks were analysed as two aspects of learning performance in this study.

4. Results

The results of the study were reported in the following order. First, total scores for each group for the whole experiment, followed by effects of sound and image match/mismatch on perceptual learning, by block-by-block improvements in terms of RA and RT on perceptual learning plus improvements in terms of tone combinations in RA for all and matched trails, and finally by effects of L1 background on learning performance in both initial and final tone sub-condition in terms of RA and RT for matched and mismatched trials were reported and compared.

4.1 Overall performance

4.1.1 Total scores for each group for the whole experiment

The total scores in terms of mean value of response accuracy (RA) for each participant for all trials were submitted to a 16 (tone combination) × 2 (L1 background) repeated measures ANOVA with tone combination as within-subjects factor, L1 background as between-subjects factor. The covariate was set up in four different alternatives: without any covariates, with musical aptitude and age as covariate separately, and with both musical aptitude and age counted as covariates. The results (listed in table 4.1) showed that both estimated marginal means of RA and degree of effect of L1 background in terms of f-value and p-value were influenced by musical aptitude counted as covariate, but negligibly little by age counted as covariate that only slightly changed f-value. Thus, it was reasonable and necessary to carry out the data analysis with musical aptitude counted as covariate in all treatment conditions in this study.

Table 4.1 Estimated marginal means (with 95% confidence interval for difference) and effect of L1 background with four set ups of covariates for both language groups. (Covariates appearing in the model were evaluated at the following values: Musical aptitude = 2,48, Age = 25,00.)

Covariate(s)	*Estimated marginal means*		*Effects*
	NS listeners	*NE listeners*	
Without	79.3%	75.3%	L1 background F (1,64) = 8.59, p = .005
Musical aptitude	78.8%	75.8%	L1 background F (1,63) = 6.22, p = .015 Musical aptitude F (1,63) = 23.20, p < .001

Age	79.3%	75.3%	L1 background F (1,63) = 8.50, p = .005 Age (1,63) = 3.68, p = .060
Musical aptitude and age	78.8%	75.8%	L1 background F (1,62) = 6.25, p = .015 Musical aptitude F (1,62) = 21.12, p < .001 Age F (1,62) = 2.30, p = .135

As expected, the total score (mean value of RA) for NS listeners (79.31%) was higher than NE listeners (75.30%). Results indicated that there was an effect of L1 background (F (1, 63) = 6.22, p = .015).

The difference in mean value of response time (RT) for all trials for NS listeners (1562.49 ms) and NE listeners (1552.66 ms) was not statistically significant.

4.1.2 Effects of sound and image match/mismatch on learning performance

4.1.2.1 RA

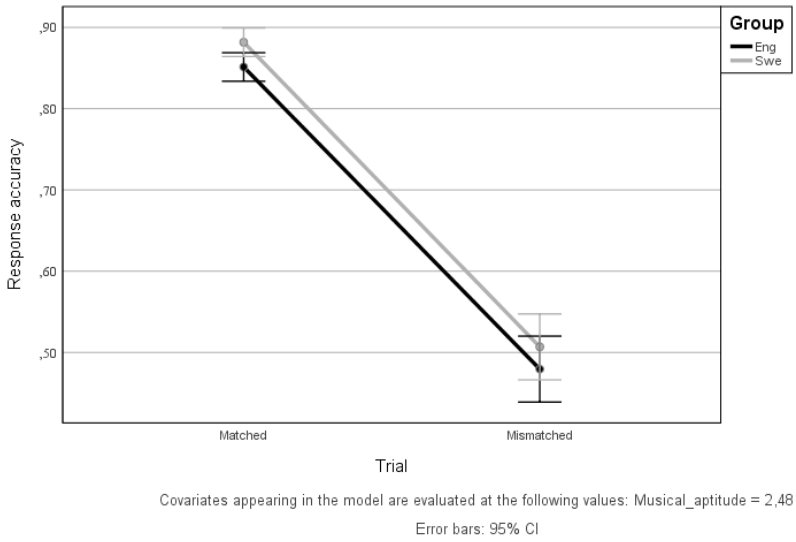
The results yielded a main effect of type of trial (F (1, 63) = 164.96, p < .001) and tone combination (F (15.945) = 2.53, p = .003). The results showed an effect of L1 background for matched trials (F (1,63) = 5.88, p = .018), but not for mismatched trials (F (1,63) = 0.89, p = .348). NS listeners and NE listeners gained a mean percentage of correct scores of 88.2% and 85.1% respectively in matched trials, and 50.7% and 48.0% respectively in mismatched trials [see figure 4.1 (a)]. The results indicated that matched trials scored significantly higher than mismatched trials for both language groups, NS listeners outperformed NE listeners both in matched trials (significantly) and mismatched trials (non-significantly), and both NS and NE listeners passed Leathers' criteria of 80-85% correct identification level for successful learning of tones in MC due to only identification was included in matched trials.

4.1.2.2 RT

The results yielded a main effect of tone combination (F (15,945) = 1.99, p = .029). Mean value of RT was 1552,67 ms for NS listeners and 1534,58 ms for NE listeners in matched trials and it was 1617,03 ms for NS listeners and 1591,88 ms for NE listeners in mismatched trials [see figure 4.1 (b)]. The mean value of RT was longer for NS listeners compared to NE listeners in both matched and mismatched trials, but the difference between the two language

groups was not statistically significant (F (1, 63) = 1.80, p = .185). The mean value of RT for matched trials was shorter than for mismatched trials for both language groups, implying that it was easier to make a judgement when the sound stimuli matched the image stimuli than when they mismatched each other for both language groups.

(a)



(b)

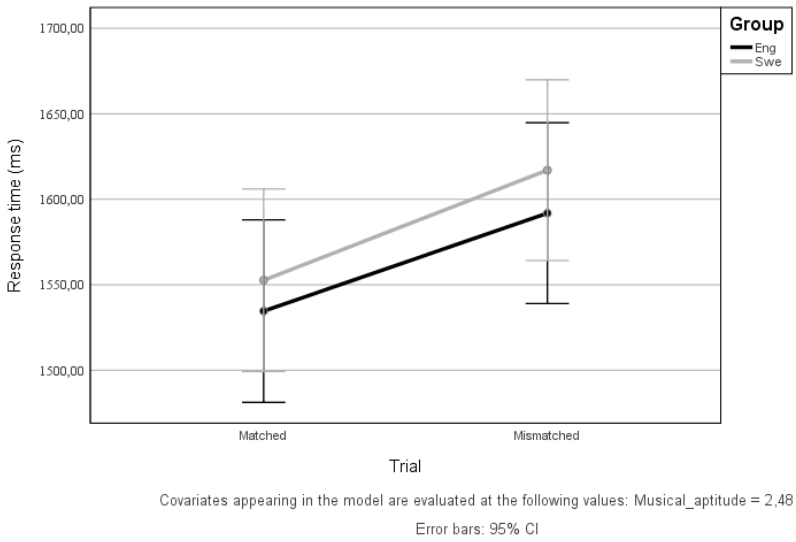


Figure 4.1 Mean scores of RA (a) and RT (b) for matched and mismatched trials for NS and NE listeners.

4.2 Block-by-block improvements on perceptual learning

4.2.1 Improvements for both groups and for each group across blocks

4.2.1.1 Improvements for all trials across blocks

4.2.1.1.1 RA

4.2.1.1.1.1 Overall improvements

The results yielded a main effect of block ($F(5,315) = 32.88$, $p < .001$), tone combination ($F(15, 945) = 2.34$, $p = .007$) and L1 background ($F(1, 63) = 6.91$, $p = .011$) for all trials, indicating that NS listeners were significantly better than NE listeners in overall performance across blocks. Analyzing the effects for each language separately, the results showed that there were effects of both block ($F(5,155) = 15.95$, $p < .001$) and tone combination ($F(15,465) = 2.45$, $p = .010$) for NS listeners and effect of block ($F(5,155) = 17.52$, $p < .001$) for NE listeners both with and without musical aptitude as covariate.

Despite significant differences in mean values of RA between NS and NE listeners, each language group made improvement block-by-block in terms of increased mean values of RA for all trials across all 6 training blocks and the largest difference in terms of mean values of RA between the two language groups was in block 3 as shown in table 4.2 (5.7 percent units difference in percentage of correct response) and in figure 4.2. The learning performance in percentage of correct response of NS listeners was 4.4% higher than NE listeners already in block 1[16]. NS listeners made greater improvements compared with NE listeners in percentage of correct response among all blocks except between block 3 and block 4, block 4 and block 5, in which NE listeners made greater progress (listed in table 4.3). NS listeners gained a substantial increase of 24.2 units in percentage of correct response (improved from 64.7% in block 1 to 88.9% in block 6) whereas NE listeners gained a substantial increase of 25.4 units in percentage of correct response (improved from 60.3% in block 1 to 85.7% in block 6) as a result of training.

[16] NS listeners were 5% higher than NE listeners in percentage of correct response in block 1 without counting musical aptitude as covariate.

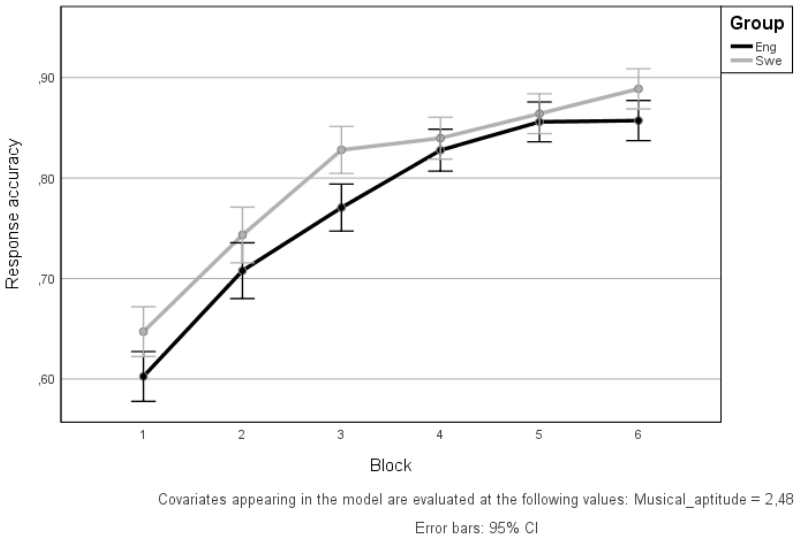


Figure 4.2 Mean values of RA for all trials across blocks for both language groups.

Table 4.2 Response accuracy in percentage across blocks for all trials for each language group (Confidence interval for difference was 95% and covariate appearing in the model were evaluated at the following values: Musical aptitude = 2.48).

	block 1	block 2	block 3	block 4	block 5	block 6
NS listeners	64.7%	74.3%	82.8%	8.0%	86.4%	88.9%
NE listeners	60.3%	70.8%	77.1%	82.8%	85.6%	85.7%

Table 4.3 Improvements in percentage of correct response between blocks for all trials for each language group.

	block 1→ 2	block 2→ 3	block 3→ 4	block 4→ 5	block 5→ 6
NS listeners	9.6	8.5	1.2	2.4	2.5
NE listeners	10.5	6.3	5.7	2.8	0.1

Table 4.4 Pairwise comparisons of estimated marginal means of RA for all trials for NS listeners (upper right triangle, read in vertical axis in terms of block) and NE listeners (lower left triangle, read in horizontal axis in terms of block). (Confidence interval for difference was 95% and covariate appearing in the model were evaluated at the following values: Musical aptitude = 2.48).

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Block 1	—	***	***	***	***	***
Block 2	***	—	*****	***	***	***
Block 3	***	*****	—	**\Ns	****	***
Block 4	***	***	**\Ns	—	Ns	Ns**
Block 5	***	***	****	Ns	—	Ns
Block 6	***	***	***	Ns**	Ns	—

Ns: not significant, *p < .05, **p < .01, ***p < .001 (Adjustment for multiple comparisons: Sidak).

Seen from pairwise comparisons of estimated marginal means of RA for all trials (listed in table 4.4), the improvement of block 3 was significant compared to block 1 (p < .001), block

2 ($p < .001$), block 6 ($p < .001$) and block 5 ($p = .026$), but not to block 4 ($p = .980$), and the improvement of block 6 was significant compared to block 1 ($p < .001$), block 2 ($p < .001$), block 3 ($p < .001$) and block 4 ($p = .001$) for NS listeners. For NE listeners, the improvement of block 3 was significant compared to block 1 ($p < .001$), block 2 ($p = .001$), block 5 ($p < .001$), block 6 ($p < .001$) and block 4 ($p = .002$), and the improvement of block 6 was significant compared to block 1 ($p < .001$), block 2 ($p < .001$) and block 3 ($p < .001$), but neither to block 4 ($p = .063$), nor to block 5 ($p = 1.000$). It can be noted that it did not differ statistically from block 5 to block 6 for NE listeners.

The results from pairwise comparison can be interpreted that NS listeners made greater improvement earlier (in block 3) than NE listeners (in block 4). Furthermore, there was no significant improvement from block 4 to block 6 for NE listeners while there was still significant improvement between block 4 and block 6 for NS listeners, thus NS listeners kept making significant progress even as late as at block 6 while NE listeners ceased making significant improvements after block 4. In sum, NS listeners were quicker learners and their space for improvements kept longer compared with NE listeners.

4.2.1.1.1.2 Improvements in terms of tone combinations

The improvement in every tone combination in terms of mean value of RA across blocks for all trials are illustrated in figure 4.3 (for NS listeners) and figure 4.4 (for NE listeners) respectively. The dynamic improvements in every tone combination in terms of mean values of RA across blocks and in terms of changes in percentage of correct response between blocks for all trials are listed in table 1 and table 2 respectively in Appendix 3. In block 1 none of the tone combination in terms of mean value of RA reached chance level (0.75) neither for NS listeners nor for NE listeners due to participants had to guess the correct match between a tone combination and the corresponding image at the very beginning. In block 2, there were seven tone combinations (T3T3, T1T1, T2T3, T2T2, T4T1, T4T2 and T3T4) for NS listeners and

only three tone combinations (T3T3, T1T1 and T3T4) for NE listeners that gained higher than chance level. In block 4, there was only one tone combination that did not surpass chance level (T4T2 for NS listeners and T4T4 for NE listeners). In block 5, all tone combinations exceeded chance level for both language groups and in block 6, all tone combinations reached a level higher than 0.8 for both language groups.

Based on the average of mean values of RA across all six training blocks, the easiest tone combinations for NS listeners for all trials were T3T3, T1T1, T2T2, T3T4, T3T2, T2T3, T3T1 and T1T3. The easiest tone combinations for NE listeners for all trials were T3T3, T1T1, T1T3, T2T2, T3T4, T3T1, T3T2 and T2T3. The most confusing tone combinations for NS listeners for all trials were T4T2, T2T4, T4T1, T1T2, T4T4, T4T3, T1T4 and T2T1. The most confusing tone combinations for NE listeners for all trials were T2T4, T4T4, T1T4, T2T1, T4T2, T4T1, T1T2 and T4T3. It can be noted that NS listeners outperformed NE listeners in all tone combinations except T4T2 for all trials. Mean values of RA for all trials for both language groups are illustrated in Figure 4.5.

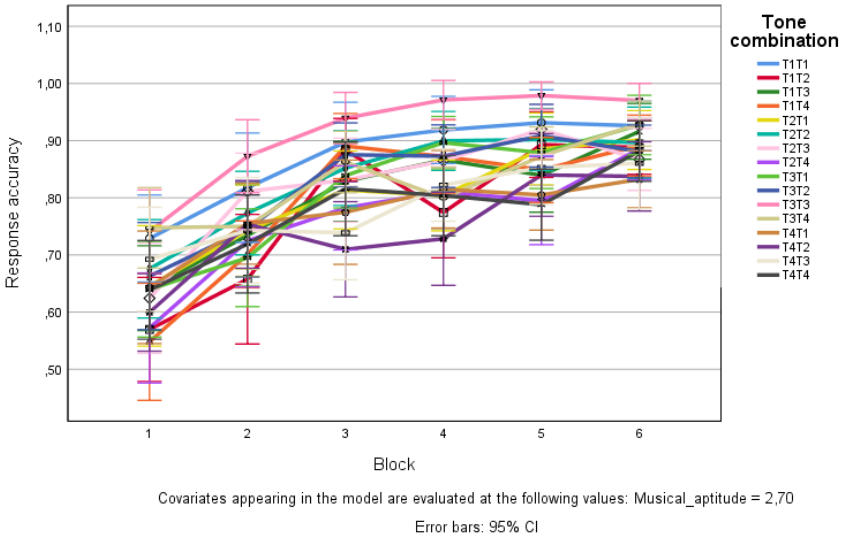


Figure 4.3 Mean values of RA of 16 tone combinations of NS listeners across blocks for all trials.

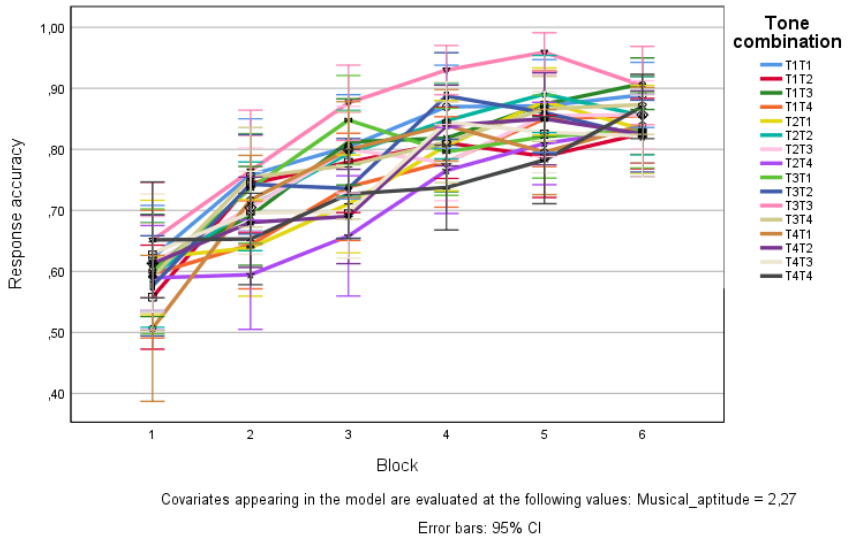
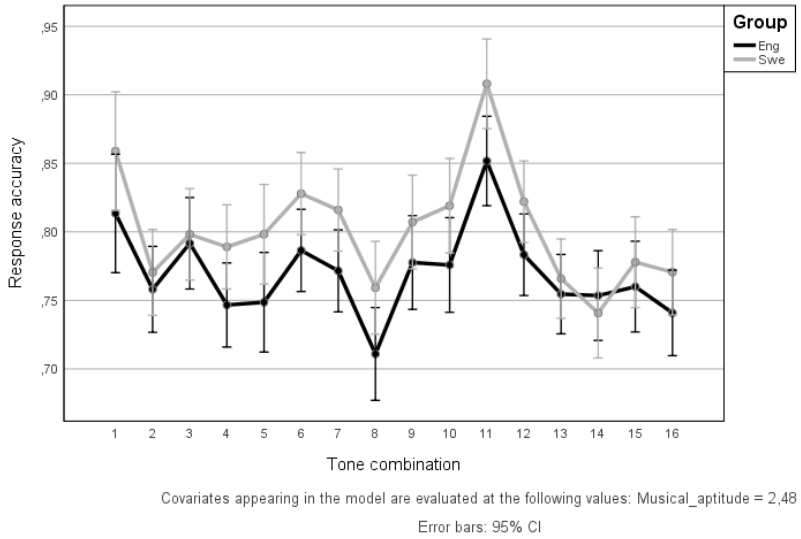


Figure 4.4 Mean values of RA of 16 tone combinations of NE listeners across blocks for all trials.



	T1T1	T1T2	T1T3	T1T4	T2T1	T2T2	T2T3	T2T4	T3T1	T3T2	T3T3	T3T4	T4T1	T4T2	T4T3	T4T4
Swe	0.859	0.770	0.798	0.789	0.798	0.828	0.816	0.759	0.807	0.819	0.908	0.822	0.766	0.741	0.778	0.771
Eng	0.813	0.758	0.792	0.747	0.749	0.786	0.772	0.711	0.778	0.776	0.852	0.783	0.755	0.753	0.760	0.741

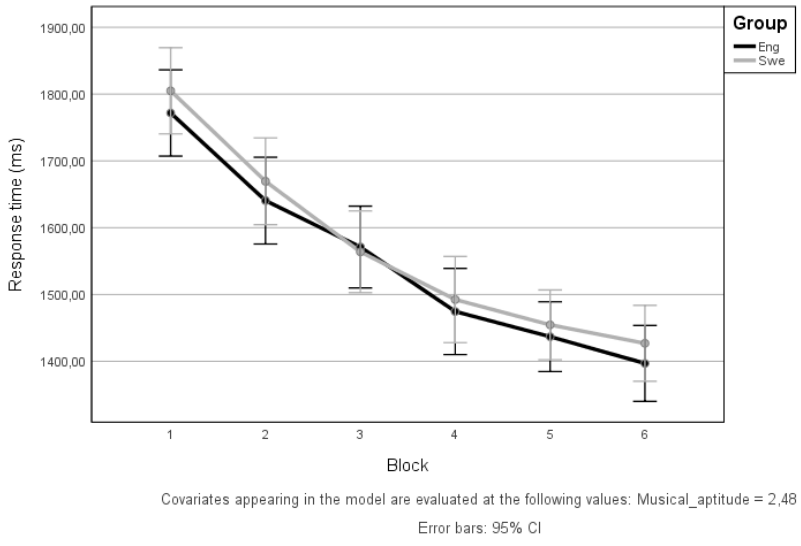
Figure 4.5 Mean values of RA of 16 tone combinations of NS and NE listeners across blocks for all trials.

4.2.1.1.2 RT

The results yielded a main effect of block ($F(5,315) = 15.20$, $p < .001$) and tone combination ($F(15,945) = 2.63$, $p = .005$). The two language groups did not differ significantly in mean value of RT for all trials across blocks ($F(1,63) = 0.285$, $p = 0.596$). The results showed an effect of block for both NS ($F(5,155) = 8.86$, $p < .001$) and NE ($F(5,155) = 6.52$, $p = .001$)

listeners, but no effect of tone combination for neither NS (F (15,465) = 1.67, p = .077) nor NE (F (15,465) = 1.38, p = .515) listeners.

The general results for all trials showed that both language groups made significant improvements block-by-block in terms of decreased mean values of RT across all blocks. The mean values of RT for NS listeners were longer than NE listeners for all training blocks with the largest difference in block 1 (a difference of 33.19 ms) and the second largest difference in block 6 (a difference of 30.04 ms), and with the exception in block 3 in which RT of NE listeners was longer instead (a difference of 7.20 ms) as shown in figure 4.6. Pairwise comparisons of estimated marginal means of RT with statistical significance among all training blocks for all trials for each language group are listed in table 4.5.



	1	2	3	4	5	6
Swe (ms)	1804.94	1669.43	1563.86	1492.45	1454.54	1426.92
Eng (ms)	1771.75	1640.54	1571.06	1474.56	1436.88	1396.88

Figure 4.6 Mean values of RT for all trials across blocks for both language groups.

Table 4.5 Pairwise comparisons of estimated marginal means of RT for all trials for NS listeners (upper right triangle, read in vertical axis in terms of block) and NE listeners (lower left triangle, read in horizontal axis in terms of block). (Confidence interval for difference was 95% and covariate appearing in the model were evaluated at the following values: Musical aptitude = 2.48).

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Block 1	—	***	***	***	***	***

Block 2	***	—	****	***	***	***
Block 3	***	****	—	****	***	***
Block 4	***	***	****	—	Ns	**
Block 5	***	***	***	Ns	—	Ns
Block 6	***	***	***	**	Ns	—

Ns: not significant, *p < .05, **p < .01, ***p < .001 (Adjustment for multiple comparisons: Sidak).

Pairwise comparisons showed that there were significant improvements in terms of decreased mean values of RT across all training blocks for all trials except that there were no significant improvements among block 4 and block 5, block 5 and block 6 for neither language group.

The mean values of RT for all trials for both language groups showed a basically contra-variant relationship with the order of blocks. It can further be noted that in block 3 when NS listeners made greater improvement in terms of high mean value of RA (illustrated in figure 4.2), they also showed an improvement in terms of decrease in mean values of RT (illustrated in figure 4.6).

4.2.1.3 Improvement for matched trials across blocks

4.2.1.3.1 RA

4.2.1.3.1.1 Overall improvements

The results yielded a main effect of block ($F(5,315) = 35.42$, $p < .001$), tone combination ($F(15, 945) = 1.80$, $p = .047$), and L1 background ($F(1, 63) = 5.89$, $p = .018$), indicating there was a block and a tone combination effect for both language groups for matched trials and NS listeners were significantly better than NE listeners in overall performance of identification of tones in MC across blocks. The results showed a block × L1 background interaction ($F(5, 315) = 2.94$, $p = .024$). Examining the effects for each language separately, the results showed an effect of block for both NS ($F(5,155) = 17.93$, $p < .001$) and NE ($F(5,155) = 18.55$, $p < .001$) listeners, but no effect of tone combination neither for NS listeners $F(15,465) = 1.79$, $p = .062$) nor for NE listeners $F(15,465) = 0.943$, $p = .490$).

Both NS and NE listeners kept learning to identify tones in MC across all training blocks and both language groups reached a level of higher than 95% in block 6. Mean values of RA in

percentage across blocks for matched trials for each language group are illustrated in figure 4.7 and listed in table 4.6, and the improvements in percentage of correct response between blocks for matched trials for each language group are listed in table 4.7.

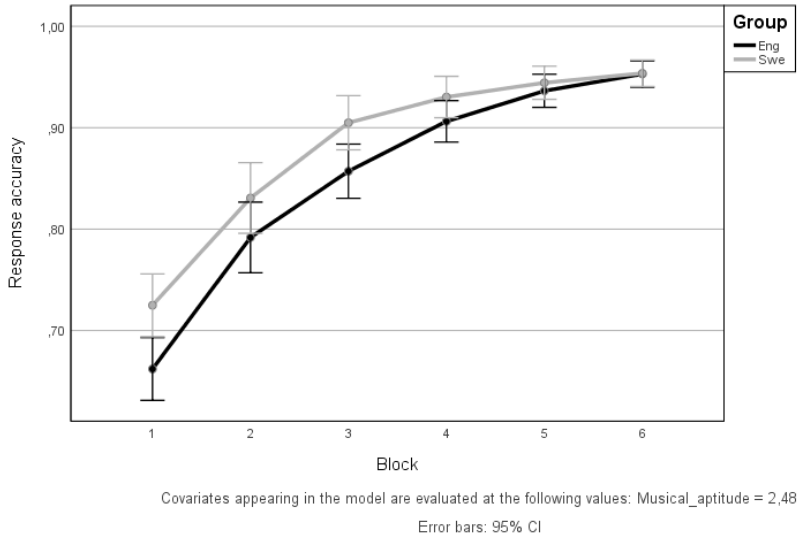


Figure 4.7 Mean values of RA for matched trials across blocks for both language groups.

Table 4.6 Response accuracy in percentage across blocks for matched trials for each language group (with 95% confidence interval for difference).

	block 1	block 2	block 3	block 4	block 5	block 6
NS listeners	72.5%	83.1%	90.5%	93.0%	94.4%	95.4%
NE listeners	66.2%	79.2%	85.7%	90.6%	93.6%	95.3%

Table 4.7 Improvements in percentage of correct response between blocks for matched trials for each language group.

	block 1→ 2	block 2→ 3	block 3→ 4	block 4→ 5	block 5→ 6
NS listeners	10.6	7.4	2.5	1.4	1
NE listeners	13	6.5	4.9	3.0	1.7

As listed in table 4.6, the identification performance in percentage of correct response for NS listeners was 6.3% higher than NE listeners already in block 1[17]. Both language groups kept making progress in terms of increase in percentage of correct response across all six training blocks. NS listeners gained an increase of 22.9 units in percentage of correct response

[17] NS listeners was 6.5% higher than NE listeners in percentage of correct identification at block 1 without counting musical aptitude as covariate.

(improved from 72.5% in block 1 to 95.4% in block 6) whereas NE listeners gained an increase of 29.1 units in percentage of correct response (improved from 66.2% in block 1 to 95.3% in block 6) as a result of training. Leather (1990) set the passing criterion for learning success at 80-85% of correct identification scores in perceptual learning of lexical tones in MC. According to this criterion, the Swedish group had learned to successfully identify the 16 tone combinations already in block 2 (83.1%) while the corresponding block for the English group was in block 3 (85.7%).

Pairwise comparisons (listed in table 4.8) showed that the improvements of block 1 and block 2 were significant compared to all other training blocks. The improvement of block 4 was only significant compared to block 1 and block 2, and the improvements of block 5 and block 6 were only significant compared to block 1, block 2 and block 3 for NS listeners while there were significant improvements among all training blocks except between block 5 and block 6 for NE listeners.

The results from pairwise comparisons can be interpreted as NS listeners having made significant improvements in identifying tones in MC already in block 3 (after block 3, the improvements were not significant anymore for NS listeners) while NE listeners kept making significant identification improvements as late as in block 5. In other words, it took less time for NS listeners to learn to identify tones in MC compared to NE listeners. Furthermore, it is worth to note that the improvement from block 5 to block 6 was not significant ($p = .991$) for matched trials for NS listeners, contrasting with the significant improvement from block 4 to block 6 ($p = .001$) for all trials for NS listeners but not for NE listeners, it is not difficult to suppose that at later blocks NS listeners made greater progress in differentiating tones in MC compared to NE listeners.

Table 4.8 Pairwise comparisons of estimated marginal means of RA for valid trials for NS listeners (upper right triangle, read in vertical axis in terms of block) and NE listeners (lower left triangle, read in horizontal axis in terms of block). (Confidence interval for difference was 95%

and covariate appearing in the model were evaluated at the following values: Musical aptitude = 2.48).

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Block 1	—	***	***	***	***	***
Block 2	***	—	*****	***	***	***
Block 3	***	*****	—	**\Ns	****	*****
Block 4	***	***	**\Ns	—	*\Ns	***\Ns
Block 5	***	***	****	*\Ns	—	Ns
Block 6	***	***	*****	***\Ns	Ns	—

Ns: not significant, *p < .05, **p < .01, ***p < .001(Adjustment for multiple comparisons: Sidak).

4.2.1.3.1.2 Improvements in terms of tone combinations

The improvements in every tone combination in terms of mean value of RA across blocks for matched trials are illustrated in figure 4.8 (for NS listeners) and figure 4.9 (for NE listeners) respectively. The dynamic improvements in every tone combination in terms of mean values of RA across blocks and in terms of changes in percentage of correct response between blocks for matched trials are listed in table 1 and 2 respectively in Appendix 4. In block 1 there were four tone combinations for NS listeners and one tone combination for NE listeners passed chance level (0.75) in terms of mean values of RA. In block 2, all tone combinations for NS listeners and 12 tone combinations for NE listeners scored higher than chance level. In block 3, all tone combinations for NS listeners and 14 tone combinations for NE listeners reached a level higher than 0.8. In block 4, all tone combinations gained a level higher than 0.8 for NE listeners. In block 6, all tone combinations exceeded a level higher than 0.9 in terms of mean values of RA for both language groups.

Based on the average of mean values of RA across all six training blocks, the easiest tone combinations for NS listeners for matched trails were T3T3, T3T4, T3T2, T2T2, T1T3, T1T1, T3T1 and T2T3. The easiest tone combinations for NE listeners for matched trails were T3T3, T3T4, T3T1, T2T2, T1T3, T1T1, T4T3 and T4T2. The most confusing tone combinations for NS listeners for matched trails were T4T2, T4T3, T2T4, T4T1, T4T4, T1T2, T2T1 and T1T4. The most confusing tone combinations for NE listeners for matched trails were T2T4, T1T4, T2T1, T4T4, T1T2, T2T3, T4T1 and T3T2. NS listeners outperformed the NE listeners in all

tone combinations except T4T2 and T4T3 for matched trails. Mean values of RA for matched trails for both language groups are illustrated in Figure 4.10.

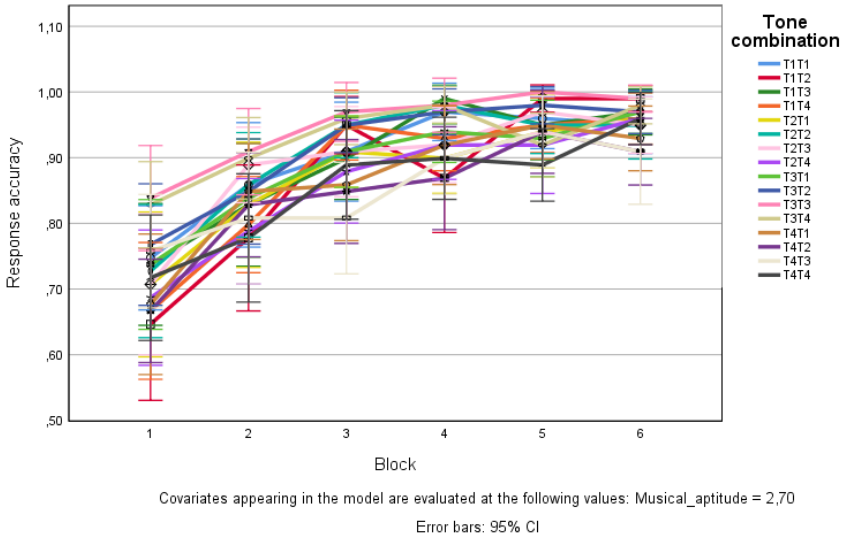


Figure 4.8 Mean values of RA of 16 tone combinations of NS listeners across blocks for matched trials.

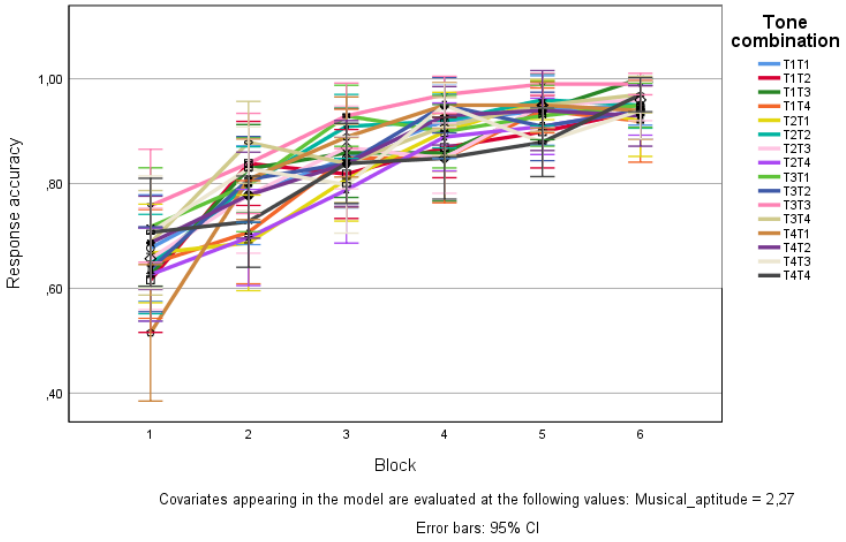
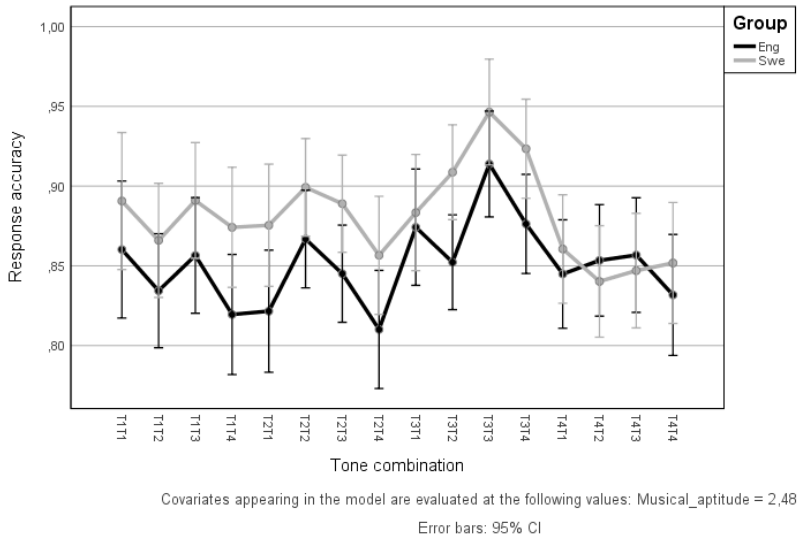


Figure 4.9 Mean values of RA of 16 tone combinations of NE listeners across blocks for matched trials.



	T1T1	T1T2	T1T3	T1T4	T2T1	T2T2	T2T3	T2T4	T3T1	T3T2	T3T3	T3T4	T4T1	T4T2	T4T3	T4T4
Swe	0.891	0.866	0.891	0.874	0.875	0.899	0.889	0.857	0.883	0.909	0.946	0.923	0.861	0.840	0.847	0.852
Eng	0.860	0.834	0.857	0.819	0.822	0.867	0.845	0.810	0.874	0.852	0.914	0.876	0.845	0.853	0.857	0.832

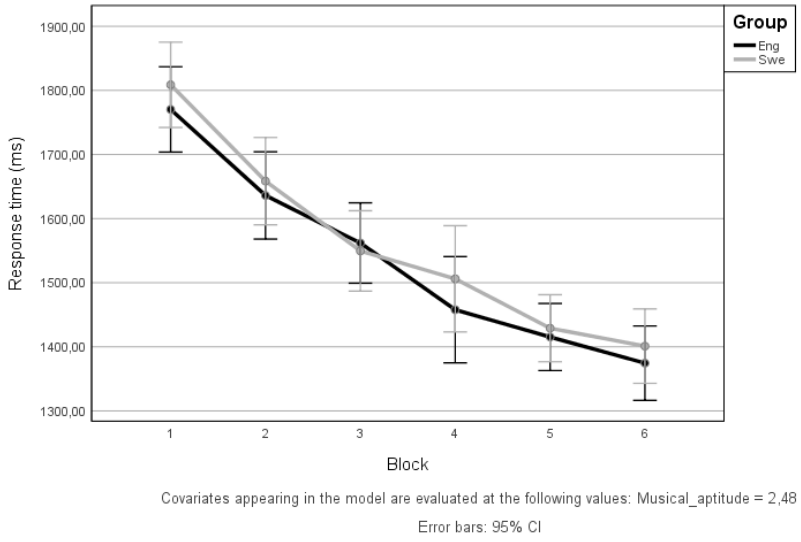
Figure 4.10 Mean values of RA of 16 tone combinations of NS and NE listeners across blocks for matched trials.

4.2.1.2.2 RT

The results yielded a main effect of block for both NS ($F(5,155) = 6.11$, $p = .001$) and NE ($F(5,155) = 6.32$, $p = .001$) listeners, but not an effect of tone combination for neither NS ($F(15,465) = 0.857$, $p = .460$) nor NE ($F(15,465) = 1.48$, $p = 0.158$) listeners. The two language groups did not differ significantly in mean value of RT for matched trials across blocks ($F(1,63) = 0.34$, $p = .562$).

The general results for matched trials showed that both language groups made significant improvements block-by-block in terms of decreased mean values of RT across all blocks. The mean values of RT for NS listeners were longer than NE listeners for all training blocks with the largest difference in block 4 (a difference of 48.24 ms) and the second largest difference in block 1 (a difference of 38.38 ms), and with the exception in block 3 in which RT of NE listeners was longer instead (a difference of 12.30 ms) as illustrated in figure 4.11. Comparing to the results of mean values of RT for all trials in which the second largest difference between NS listeners and NE listeners was in block 6 (see 4.2.1.1.2) and the result of mean values of RA in block 4 was only significant to block1 and block 2 for NS listeners for

matched trials, it can be interpreted that the longer RT for NS listeners in block 6 for all trials was due to more focus on discriminating tones in MC instead of identifying them and this focus began as early as in block 4.



	1	2	3	4	5	6
Swe (ms)	1770,28	1636,11	1561,89	1457,82	1415,29	1374,49
Eng (ms)	1808,66	1658,42	1549,60	1506,06	1429,00	1401,01

Figure 4.11 Mean values of RT for matched trials across blocks for both language groups.

As listed in table 4.9, the results of pairwise comparisons showed that the improvement of block 5 ($p < .001$) was significant compared to block 1, block 2 and block 3 for both NS and NE listeners, the improvements of block 1 ($p < .001$) and block 2 ($p = .049$ to block 4 and $p < .001$ to other blocks) were significant compared to all other training blocks for NS listeners whereas the improvements of block 1 ($p < .001$), block 2 ($p = .020$ to block 3 and $p < .001$ to other blocks) and block 3 ($p = .020$ to block 2 and $p < .001$ to other blocks) were significant compared to all other training blocks for NE listeners, the improvement of block 4 was significant compared to block 1 ($p < .001$) and block 2 ($p = .049$) for NS listeners and was significant compared to block 1 ($p < .001$), block 2 ($p < .001$), block 3 ($p < .001$) and block 6 ($p = .001$) for NE listeners, and the improvement of block 6 was significant compared to block 1 ($p < .001$), block 2 ($p < .001$) and block 3 ($p < .001$) for NS listeners and was significant compared to block 1 ($p < .001$), block 2 ($p < .001$), block 3 ($p < .001$) and block 4

(p = .001) for NE listeners. The results can be interpreted such that the significant decrease of mean value of RT ceased in block 3 for NS listeners but in block 5 for NE listeners. In other words, the RT for identifying tones in MC was relatively stable in block 3 for NS listeners and in block 5 for NE listeners. As in all trials, the results of mean values of RT for NS and NE listeners for matched trials basically showed a contra-variant relationship with the order of blocks.

Table 4.9 Pairwise comparisons of estimated marginal means of RT for valid trials for NS listeners (upper right triangle, read in vertical axis in terms of block) and NE listeners (lower left triangle, read in horizontal axis in terms of block). (Confidence interval for difference was 95% and covariate appearing in the model were evaluated at the following values: Musical aptitude = 2.48).

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Block 1	—	***	***	***	***	***
Block 2	***	—	***	****	***	***
Block 3	***	****	—	***\Ns	***	***
Block 4	***	****	***\Ns	—	Ns	**\Ns
Block 5	***	***	***	Ns	—	Ns
Block 6	***	***	***	**\Ns	Ns	—

Ns: not significant, *p < .05, **p < .01, ***p < .001 (Adjustment for multiple comparisons: Sidak).

4.2.2 Summary of improvements of RA and RT for all and matched trials across blocks

Both language groups made improvements in terms of increased mean values of RA and decreased mean values of RT across blocks for both all and matched trials. NS listeners scored significantly higher than NE listeners in mean values of RA across blocks for both all and matched trials. There was no significant difference in mean values of RT between the two language groups though the mean values of RT were longer for NS listeners than for NE listeners in all blocks with an exception for block 3 for both all and matched trials.

Concerning improvements in terms of percentage of correct response, NS listeners gained a substantial increase of 24.2 percent units in all trials and 22.9 percent units in matched trials and the corresponding increase for NE listeners was 25.4 percent units in all trials and 29.1 percent units in matched trials as a result of 6 blocks' perceptual training. NS listeners scored

higher than NE listeners already in block 1 for matched trials (6.3%) and for all trials (4.4%), and pairwise comparisons showed that NS listeners ceased to make significant improvement in block 3 for matched trials but kept making significant improvement even in block 6 for all trials while NE listeners ceased to make significant improvement in block 5 for matched trials and in block 4 for all trials. Thus, it can be implied from the pairwise comparisons that NS listeners are quicker learners both in identifying (from block 1 and block 3) and discriminating (from block 4 to block 6) tones in MC compared with NE listeners. In addition, for matched trials, the result of mean values of RA in block 4 was only significant to block 1 and block 2 for NS listeners and the largest difference of mean values of RT was in block 4 between NS and NE listeners, further implying that NS listeners began learning to discriminate tones in MC already in block 4 and kept learning until the final training block but this learning ability of discrimination was missing for NE listeners because NE listeners made significant improvements until block 5 for matched trials and they ceased to make significant improvements in block 4 for all trials.

Regarding to influence of musical aptitude on perceptual learning of tones in MC, the effect of block, tone combination and L1 background became weaker with musical aptitude as covariate compared to without in terms of mean values of RA for both all trials and matched trials. Examining the influence of musical aptitude on each language group separately, it was found that the effect of tone combination on mean values of RA disappeared for NE listeners for all trials and for both language groups for matched trials with musical aptitude as covariate compared to without, and the effect of musical aptitude on mean values of RA was greater for NS listeners than NE listeners for both all trials and matched trials. In terms of influence of musical aptitude on mean values of RT, the effect of block became slightly weaker for all trials and matched trials and the effect of tone combination disappeared for all trials and matched trials for both language groups with musical aptitude as covariate compared to

without. The difference between NS and NE listeners in terms of RT became slightly greater with musical aptitude as covariate compared to without. There was only effect of musical aptitude for NS listeners but not for NE listeners on mean values of RT in all trials and matched trials.

The results revealed that every participant made progress across blocks and every tone combination was better learned across blocks in general though little decrease occurred among some blocks for some of the tone combinations, but the whole trend was kept learning and improving. It should also be noted that there was a large degree of variability among participants in both language groups.

Because the main purpose of this study was to investigate the ability of associating different tone combinations to corresponding lexical meanings and because the presentation of stimuli was counterbalanced across all training blocks and participants, the 16 tone combinations with mismatched images were not evenly distributed in every training block. Thus, mean values of RA and RT in mismatched trials across blocks were not compiled in this study.

4.3 Effects of L1 background on learning performance

Results of effects of L1 background were reported in both initial and final tone sub-condition in terms of mean values of RA and RT for matched and mismatched trials in order to better compare them with the processing of pitch patterns of accent 1 and accent 2 in Swedish. Mean values of RA and RT for matched and mismatched trials for both NS and NE listeners in both initial and final tone sub-condition are illustrated in figure 1 in Appendix 5 and effects of pairwise comparisons of RA and RT in both initial and final tone sub-condition for matched and mismatched trials for both NS and NE listeners are listed in table 1 and 2 respectively in Appendix 5.

4.3.1 Effects of L1 background on learning performance for matched trials

4.3.1.1 RA

The results indicated an effect of initial tone F (3,93) = 4.02, p = .011) for matched trials for NS listeners and an effect of L1 background F (1,63) = 5.88, p = .018).

Pairwise comparisons of the same initial tone for matched trials showed that mean value of RA of T4T* was significantly lower than mean value RA of T3T* (p < .001) and mean value of RA of T3T* was significantly higher than mean value RA of all other tones [T1T*(p = .016), T2T*(p = .001), T4T*(p < .001)] for NS listeners whereas mean value RA of T3T* was significantly higher than mean value RA of T1T* (p = .008) and T2T* (p = .006) and mean value RA of T4T* did not differ significantly from mean value RA of any other tones as initial tone for NE listeners.

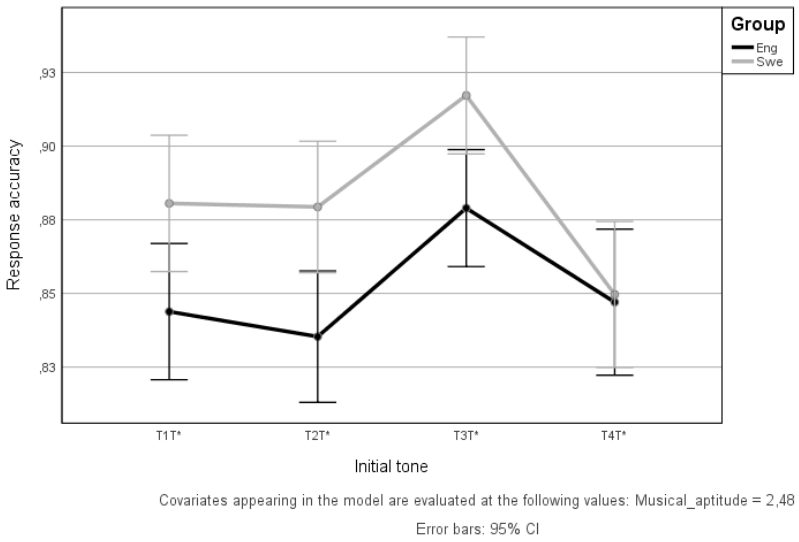
On the other hand, it can be noted from pairwise comparisons of the same initial tone for matched trials that mean value of RA of T1T* did not differ statistically from T2T* (p = .999) for neither NS nor NE listeners whereas mean value of RA of T1T* did not differ statistically from T4T* (p = 1.000) for NE listeners.

Pairwise comparisons of the same final tone for matched trials indicated that mean value RA of T4 as final tone was significantly lower than mean value RA of T3 (p = .008) as final tone for NE listeners while there were no significant differences among mean values RA of all other tones as final tone for NS listeners. It can also be noted from pairwise comparisons of the same final tone for matched trials that mean value of T*T1 and T*T2 (p = 1.000) did not differ statistically for both NS and NE listeners, and mean value of T*T1 and T*T4 (p = 1.000) did not differ statistically and T*T2 did not differ statistically from T*T4 (p = 0.997) for NE listeners.

Thus, it can be interpreted that only as initial tone, T4 differed significantly in mean value of RA from T3 in the same position, and as final tone T4 did not differ significantly from any

other tones in the same position for matched trials for NS listeners. For NE listeners, the result was reversed. In other words, the mean value of RA of T4 as initial tone did not differ significantly from any other tones in the same position and mean value RA of T4 as final tone was significantly lower than mean value RA of T3 as final tone for NE listeners. Mean values of RA of initial (a) and final (b) tone sub-condition for matched trials for both language groups are illustrated in figure 4.12.

(a)



(b)

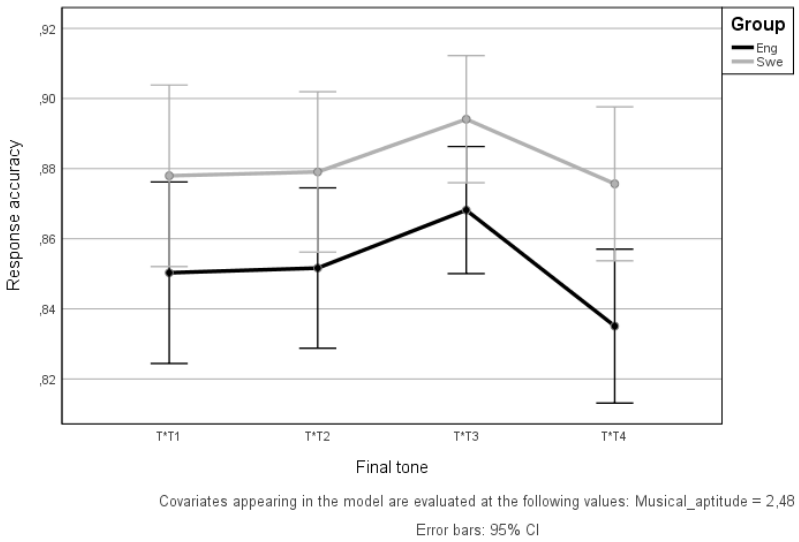


Figure 4.12 Mean values of RA of initial (a) and final (b) tone sub-condition for matched trials for both language groups.

4.3.1.2 RT

The results indicated an effect of initial tone F (3,93) = 3.13, p = .037) for matched trials for NS listeners. The results yielded no effect of L1 background (F (1,63) = 0.226, p = .636).

Pairwise comparisons of the same initial tone for matched trials showed that mean value of RT of T4T* was significantly longer than mean value of RT of all other three tones [T1T*(p < .001), T2T*(p < .001), T3T*(p < .001)] as initial tone and mean value of RT of T3T* was significantly shorter than mean value of RT of all other tones [T1T*(p = .005), T2T*(p < .001), T4T*(p < .001)] as initial tone for NS listeners and for NE listeners, mean value of RT of T4T* was significantly longer than mean value of RT of T3T* (p < .001) and mean value of RT of T3T* was significantly shorter than mean value of RT of all other tones [T1T*(p = .029), T2T*(p = .003), T4T* (p < .001)] as initial tone. It can be noted that T4T* did not differ statistically from T2T* (p = .914) for NE listeners.

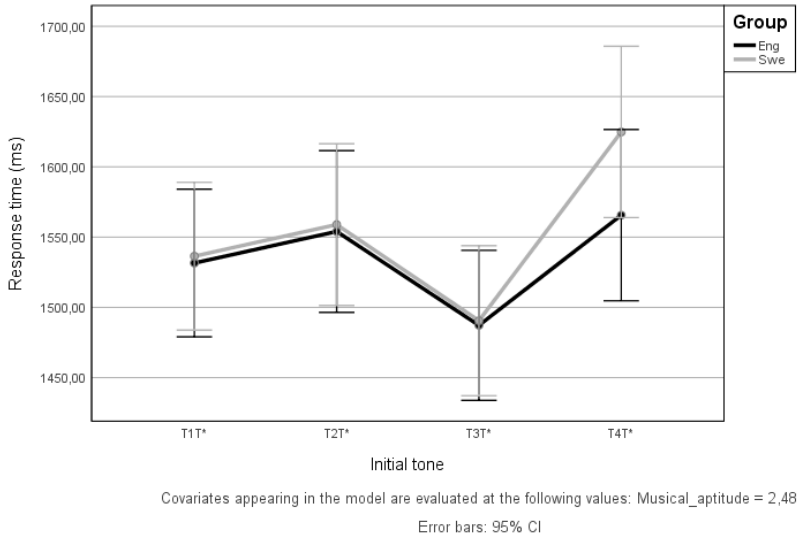
Pairwise comparisons of the same final tone for matched trials indicated that mean value of RT of T4 as final tone did not differ significantly from mean value of RT of all other three tones as final tone for NS listeners and mean value of RT of T*T4 was significantly shorter than mean value of RT of T*T2 (p < .001) for NE listeners whereas mean value of T3 as final tone was significantly longer than mean value of RT of T*T1 for both NS listeners (p = .020) and NE listeners (p = .024). It can be noted that the time to identify T2 as final tone was longest among all other tones in the same position for both NS and NE listeners.

Thus, to summarize the results of mean value of RT of initial and final tone sub-condition for matched trials, mean value of RT of T4T* was significantly longer than any other tones as initial tone and mean value of RT of T*T4 did not differ statistically from any other tones as final tone for NS listeners while mean value of RT of T4T* was significantly longer than

mean value of RT of T3T* and mean value of RT of T*T4 was significant shorter than mean value of RT of T*T2 for NE listeners.

Furthermore, it did not differ statistically in mean value of RT for NE listeners in identifying T4T* and T2T* and the time to identify T*T2 was longest among all other tones in the same position for both NS (not statistically significant) and NE listeners (statistically significant). Mean values of RT of initial (a) and final (b) tone sub-condition for matched trials for both language groups are illustrated in figure 4.13.

(a)



(b)

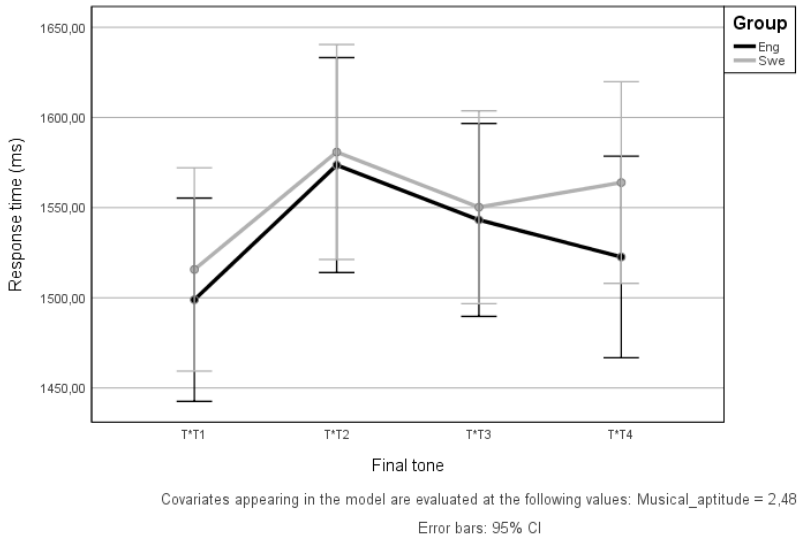


Figure 4.13 Mean values of RT of initial (a) and final (b) tone sub-condition for matched trials for both language groups.

4.3.2 Effects of L1 background on learning performance for mismatched trials

4.3.2.1 RA

The results indicated no effect of neither initial nor final tone for neither NS nor NE listeners in mean values of RA for mismatched trials.

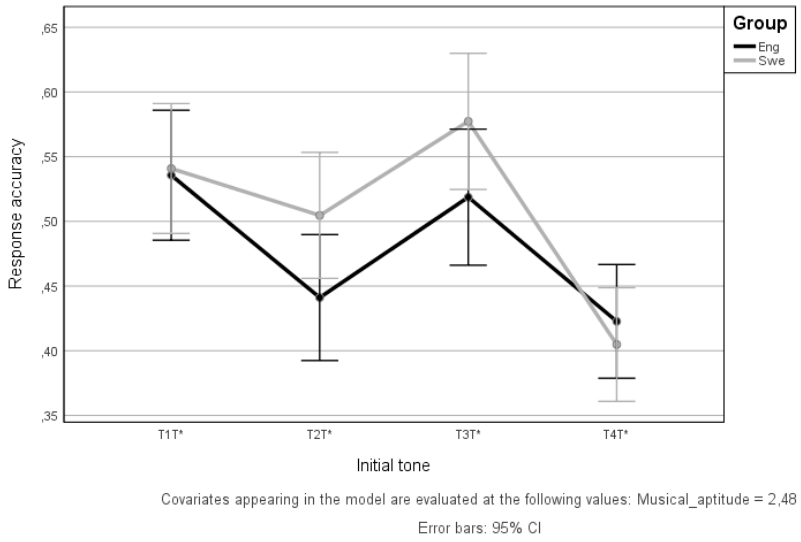
Pairwise comparisons of the same initial tone for mismatched trials showed that mean value of RA of T4T* was significantly lower than mean value of RA of T1T* ($p < .001$), T2T* ($p < .001$) and T3T* ($p < .001$) for NS listeners and mean value of RA of T4T* was significantly lower than mean value of RA of T1T* ($p = .001$) and T3T* ($p = .003$) for NE listeners whereas mean value of RA of T3T* was significantly higher than mean value of RA of T2T* ($p = .008$ for NS listeners and $p = .004$ for NE listeners) and T4T* ($p < .001$ for NS listeners and $p = .003$ for NE listeners) for both language groups. It worth to note from pairwise comparisons of the same initial tone that mean value of RA of T2T* did not differ statistically to mean value of RA of T4T* ($p = .995$) for NE listeners.

Pairwise comparisons of the same final tone for mismatched trials indicated that mean value of RA of T*T4 was significantly lower than mean value of RA of T*T1 ($p = .009$) and T*T3 ($p = .002$) for NS listeners and mean value of RA of T*T4 was significantly lower than mean

value of RA of T*T3 (p < .001) for NE listeners while mean value of RA of T*T3 was significantly higher than mean value of RA of T*T2 (p = .002 for NS listeners and p < .001 for NE listeners) and T*T4 (p = .002 for NS listeners and p < .001for NE listeners) for both language groups. It is also worth to note from pairwise comparison of the same final tone that mean value of RA of T*T2 did not differ statistically from T*T4 for NS listeners (p = 0.998) and mean value of RA of T*T2 and T*T4 did not differ statistically (p = 1.000) for NE listeners.

Thus and so, it can be interpreted that, for mismatched trials, T4T* was most difficult for NS listeners to differentiate from all the other tones as initial tone while T4T* was significantly more difficult than T1T* and T3T* for NE listeners, and T*T4 was significantly more difficult than T*T3 for both NS and NE listeners. Mean values of RA of initial (a) and final (b) tone sub-condition for mismatched trials for both language groups are illustrated in figure 4.14.

(a)



(b)

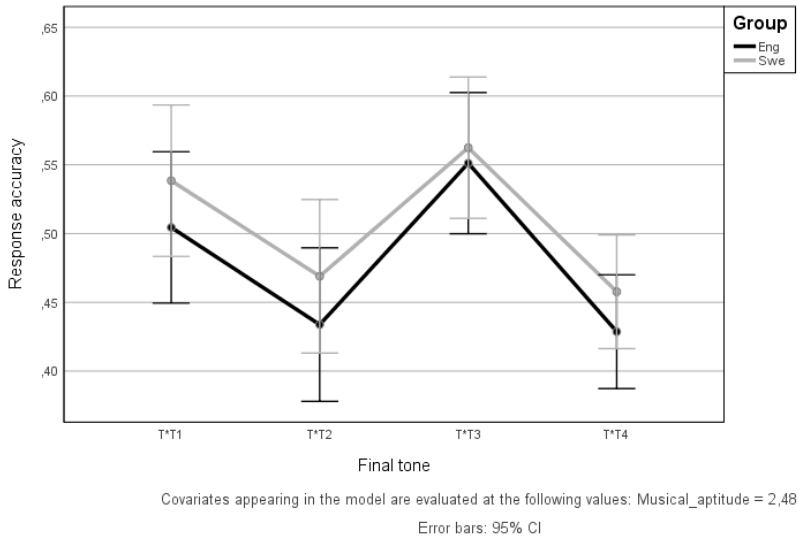


Figure 4.14 Mean values of RA of initial (a) and final (b) tone sub-condition for mismatched trials for both language groups.

4.3.2.2 RT

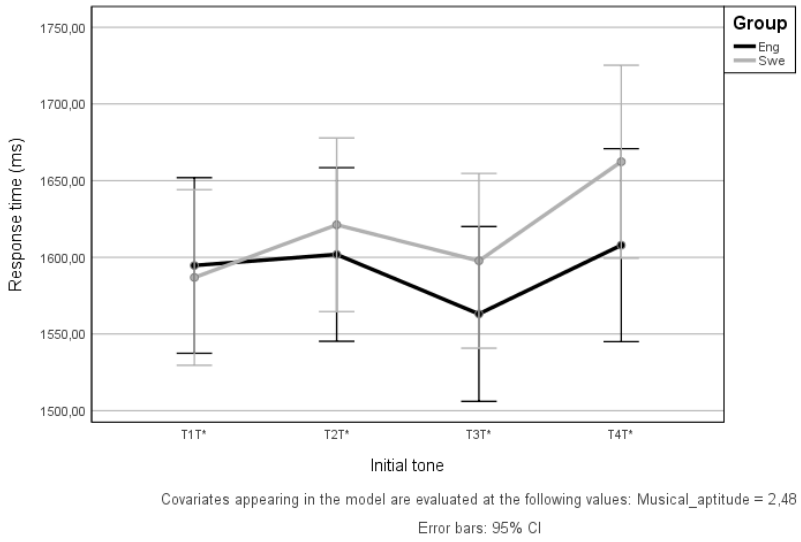
The results indicated no effect of initial tone, final tone, L1 background, musical aptitude for either language groups. Pairwise comparison of the same initial tone for mismatched trials showed that mean value of RT of T4T* was significantly longer than mean RT of T1T*(p = .001) and T3T* (p = .048) for NS listeners and there were no significant differences in initial tone sub-condition for NE listeners. It is worth to note from pairwise comparisons of the same initial tone that mean value of RT of T1T* and T2T* did not differ statistically (p = 1.000) nor did the mean value of RT of T1T* differ from that of T4T*(p = .999), or the mean value of RT of T2T* from that of T4T*(p = .998) for NE listeners.

Pairwise comparisons of the same final tone for mismatched trials indicated that there were no significant differences in final tone sub-condition for neither NS listeners nor NE listeners. However, it is worth to note from pairwise comparison of the same final tone that mean value of RT of T*T4 and T*T2 did not differ statistically (p = .984) for NS listeners and mean value of RT of T*T1 and T*T4 (p = 1.000) did not differ statistically for NE listeners.

Thus, it can be summarized that, for mismatched trials, only mean RT of T4T* was significantly longer than mean RT of T1T* and T3T* for NS listeners while there were no

significant differences in initial tone sub-condition for NE listeners and in final tone sub-condition for neither NS nor NE listeners. Mean values of RT of initial (a) and final (b) tone sub-condition for mismatched trials for both language groups are illustrated in figure 4.15.

(a)



(b)

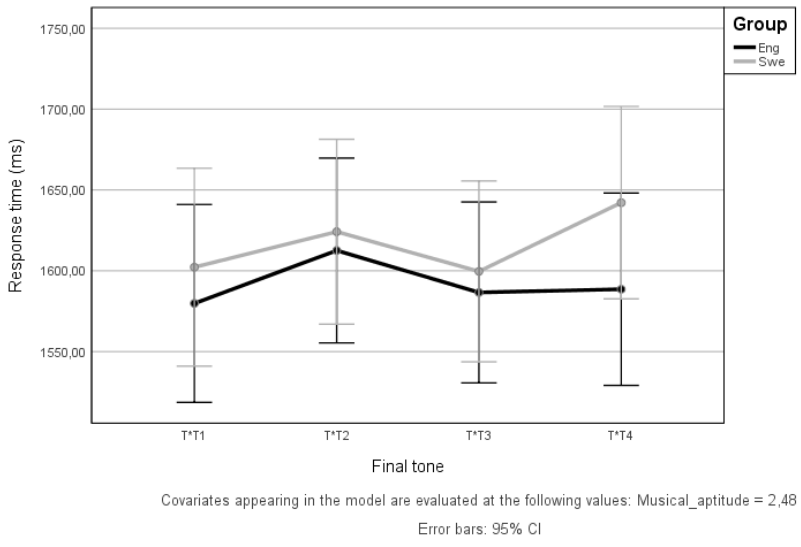


Figure 4.15 Mean values of RT of initial (a) and final (b) tone sub-condition for mismatched trials for both language groups.

4.3.3 Summary of effects of L1 background on learning performance for matched and mismatched trials

4.3.3.1 Effect of initial tone

Overall and in terms of mean values of RA, the results showed an effect of initial tone for matched trials for NS listeners and neither effect of initial nor final tone for mismatched trials for neither language groups. There was an effect of L1 background for matched trials, but not for mismatched trials.

The results indicated that T4 as initial tone (T4T*) was significantly more difficult than T3 as initial tone (T3T*) in matched trials and was significantly more difficult than all other tones as initial tone in mismatched trials for NS listeners. Thus, NS listeners had only difficulty to identify whether it was T4 or T3 as initial tone but had difficulty to discriminate T4 as initial tone from all other tones in the same position. For NE listeners, T4 as initial tone did not differ significantly from any other tones as initial tone in matched trials and T4 as initial tone was significantly more difficult than T1 and T3 as initial tone in mismatched trials. Thus, NE listeners had significantly more difficulty to differentiate T4 as initial tone from T1 and T3 in the same position and T4 as initial tone was not significantly more difficult to identify compared to all other tones in the same position for NE listeners.

For NS listeners, T4 as final tone was not significantly more difficult to identify compared to other tones as final tone, but T4 as final tone was significantly more difficult to discriminate than T1 and T3 as final tone. For NE listeners, T4 as final tone was significantly more difficult both to identify and to discriminate than T3 as final tone.

Overall and in terms of mean values of RT, the results showed an effect of initial tone for matched trials for NS listeners and neither effect of initial nor final tone for mismatched trials for neither language groups. There was no effect of L1 background for neither matched nor mismatched trials.

The results showed that mean value of RT of T4 as initial tone was significantly longer than any other tones as initial tone in matched trials and was significantly longer than T1 and T3 as initial tone in mismatched trials for NS listeners. Thus, it took significantly longer time for NS listeners to identify T4 as initial tone than other tones as initial tone and it took significantly longer time for NS listeners to discriminate T4 as initial tone than T1 and T3 as initial tone. For NE listeners, mean value of RT of T4 as initial tone was significantly longer than T3 as initial tone in matched trials and RT of T4 as initial tone was not significantly longer than any other tones as initial tone in mismatched trials. Thus, it took significantly longer time to identify T4 as initial tone than T3 as initial tone and there was no significant difference in mean value of RT to discriminate different tones as initial tone for NE listeners.

For NS listeners, there was no significant difference in mean value of RT of T4 as final tone neither in matched trials nor in mismatched trials. For NE listeners, there was no significant difference in mean value of RT of T4 as final tone in mismatched trials while it took significantly shorter time to identify T4 as final tone compared to T2 as final tone.

4.3.3.2 Tone combinations differed non-statistically

The results showed that it did not differ statistically in terms of RA and RT between tone combinations such as T1T* & T2T* vs. T*T1 & T*T2, T1T* & T4T* vs. T*T1 & T*T4 and T2T* & T4T* vs. T*T4 & T*T2 for both language groups, though it differed to some extent between the language groups owing to the tone sequence of tone combinations. The tone combinations differed non-statistically were summarized as follows:

It did not differ statistically in correct identification between T1 and T2 neither as initial and nor as final tone for neither NS nor NE listeners, and T1 and T2 as initial tone did not differ statistically in discrimination for NE listeners in terms of RT.

It differed non-statistically in identification accuracy between T1 and T4 as initial tone for NE listeners and between T1 and T4 as final tone for NS listeners, and T1 and T4 as final tone differed non-statistically in discrimination for NE listeners in terms of RT.

In addition, it did not differ statistically in correct identification between T2 and T4 as initial tone for NE listeners and as final tone for NS listeners, and it did not differ statistically in correct discrimination between T2 and T4 as initial tone for NE listeners and as final tone for both NS and NE listeners. In terms of RT, it did not differ statistically between T2 and T4 as initial tone in identification and discrimination for NE listeners and it did not differ statistically between T2 and T4 as final tone in discrimination for NS listeners.

4.3.3.3 Influence of musical aptitude on effect of initial and final tone

Regarding influence of musical aptitude on mean value of RA in terms of effect of initial tone and final tone, the effect of initial tone for matched trials for NE listeners and the effect of both initial and final tone for mismatched trials for both language groups disappeared with musical aptitude as covariate compared to without, only the effect of initial tone for matched trials for NS listeners remained. Effect of L1 background for matched trials became weaker with musical aptitude as covariate compared to without. The effect of musical aptitude on mean values of RA was greater for NS listeners than NE listeners for matched trials. For mismatched trials, there was only an effect of musical aptitude and an interaction of initial tone × musical aptitude for NS listeners, but not for NE listeners.

Regarding influence of musical aptitude on mean value of RT in terms of effect of initial and final tone, the effect of initial tone for matched trials for NE listeners and the effect of both initial and final tone for mismatched trials for both language groups disappeared with musical aptitude as covariate compared to without, only the effect of initial tone for matched trials for NS listeners remained. Opposite to influence of musical aptitude on mean value of RA, effect of L1 background (not significant) in terms of RT for matched trials was slightly greater with

musical aptitude as covariate compared to without. The results showed an effect of musical aptitude and an interaction of initial tone × musical aptitude for NS listeners for matched trials, but there was no effect of musical aptitude for NE listeners for matched trials and for neither language groups for mismatched trials.

5. Discussion

As mentioned in both the introduction and background parts of this study, the behavioural experiment of this study was designed to answer the question of whether native listeners of a pitch accent language (Swedish) can outmatch native listeners of a non-lexical-tone and non-pitch-accent language (English) in learning to perceive tones from a lexical tone language (MC) with which they have no previous experience. Furthermore, this study sought to assess whether the auditory-image (AI) training paradigm would be more effective in enhancing the learnability of lexical tones at a naïve level in a short-time laboratory training.

Comparing this study with previous perceptual studies in both tone and pitch accent, it is found that the results of this study extend the previous findings (Roll, et al., 2010, 2011, 2013, 2015, 2017; Söderström. et al., 2012,2016; Schremm, et al. 2016) indicating that NS listeners intrinsically use word stem tones to rapidly predict and pre-activate upcoming suffixes – in demonstrating that this perceptual cue for NS listeners is transferable in perceiving lexical tones in MC. Moreover, this study adds to previous findings (Chandrasekaran et al., 2010, Wong, et al. 2007) confirming that the phonetic–phonological–lexical continuity training paradigm is effective in increasing non-native listeners' phonological awareness and general auditory ability in perceiving lexical tones in MC.

5.1 Effects of L1 background on learning performance

5.1.1 Initial and final tone sub-condition

5.1.1.1 Initial and final tone sub-condition for matched trials

There was an effect of L1 background on identification accuracy of tones in MC and this is in accordance with the hypothesis that cross-linguistic perception of lexical tones in MC is influenced in large part by native linguistic experience of L1 (Best, 1995). In terms of RA, the results yielded an effect of initial tone for NS listeners, but not for NE listeners. To examine the intriguing results more closely, T4T* had significantly lower identification accuracy than

T3T*, and T*T4 did not differ significantly from any other tones as final tone in identification accuracy for NS listeners. If it can be speculated that T3T* triggers an accent 1 identification due to a low initial stem tone and T4T* triggers an accent 2 identification due to a high initial stem tone for NS listeners as predicted in 2.6.1, this result is in partly agreement with Felder et al. (2009) in which the identification accuracy of accent 1 words was higher (but not significantly) than accent 2 words. The results were reversed for NE listeners in that the mean value of RA of T4T* did not differ significantly from any other tones as initial tone and mean value of RA of T*T4 was significantly lower than mean value RA of T*T3. Thus, the effect of initial stem tone of the falling pitch is missing for NE listeners in the present study.

Interestingly, seen from the perspective of another pair of tone combinations, the results showed that T1T* did not differ statistically from T2T* ($p = .999$) in identification accuracy for neither NS nor NE listeners. As mentioned in section 2.2.1.1, T1 (55) and T2 (35) share similar pitch offset but differ in pitch onset (initial tone). However, this result showed no effect of initial tone in perceiving T1T* and T2T* for NS listeners and the absence of this effect may be speculated to be due to the reason that, unlike T3T* and T4T* in MC having an initial falling pitch that make them comparable to the pitch falls of accent 1 and accent 2 in Swedish, T1T* and T2T* in MC have rather an even and a rising initial pitch contour respectively and thus the processing of accent 1 and accent 2 in Swedish cannot be triggered in identifying T1T* and T2T* for NS listeners. On the other hand, the identification accuracy of T1T* and T2T* differing non-significantly for both NS and NE listeners may be explained by the fact that neither a high even (T1) nor a middle rising (T2) initial tone exists neither in words in isolation nor in statements (in the sense of sentence type) neither in Swedish nor in English and consequently, the linguistic experience in perceiving the differences between a high even (T1) or a middle rising (T2) initial tone is equally minimal for both NS and NE listeners. Moreover, for both NS and NE listeners, the results also showed that T*T1 and

T*T2 (p = 1.000) did not differ statistically in identification accuracy either. Thus, it is reasonable to presume, that both NS and NE listeners might rely on other acoustic cues, for instance intensity contour (shown in figure 2.3), instead of pitch contour in perceiving T1T* and T2T* (or T*T1 and T*T2) due to T1 and T2 sharing a similar amplitude contour in stimuli (shown in figure 2.2). Furthermore, the result showed that mean value of RA of T1T* did not differ statistically from T4T* (p = 1.000) in identification accuracy for NE listeners and this result is in agreement with Lehiste (1970), White (1981) and Shen (1989), thus it affords more information for the speculation that NE listeners tended to perceive high pitch onset on T1 and T4 in MC as stressed syllable in English and thus were confused with these two tones in MC. For NS listeners, this aspect of the results went reversed, namely, T*T1 and T*T4 (p = 1.000) did not differ statistically in identification accuracy and it might be explained that NS listeners are also more sensitive to the final stem tone (T1 and T4 share have similar pitch onset) instead of the final pitch contour. What makes it difficult to explain is the result that T2T* did not differ significantly from T4T* (p = 0.981) in identification accuracy for NE listeners and T*T4 did not differ significantly from T*T2 (p = 0.997) in identification accuracy for NS listeners. One explanation – of a highly speculative nature – could be that the "ΔF/duration" (the velocity of the pitch fall, see 2.2.4) in T4 is so fast compared to the pitch fall in accent 2 words in Swedish that it was misidentified as a rising tone instead[18]. Nevertheless, this speculation cannot explain why NE listeners have the similar difficulty in identifying T2T* and T4T*.

In terms of RT, the results yielded an effect of initial tone for NS listeners, but not for NE listeners and there was no significant difference between NS and NE listeners in response time of identifying tones in MC in general. Analysing pairwise comparison of initial tone sub-

[18]After the experiment, one participant told the experimenter that the quick fall of T4 in MC made her feel dizzy and consequently she lost her judgement, not unlike the feeling of riding down hills.

condition for each language group separately and concretely, the results showed that mean value of RT of T4T* was significantly longer than any other tones as initial tone and mean value of RT of T*T4 did not differ statistically from any other tones as final tone for NS listeners. The significantly longer response time in identifying T4T* for NS listeners is in accordance with Söderström et al. (2012) in that the response time for accent 2 suffix following a correctly associated accent 2 stem tone was longer than accent 1 suffix following a correctly associated accent 1 stem tone, and with Felder et al. (2009) in that it took significantly longer time to identify accent 2 words compared to accent 1 words. However, the results showed that the mean value of RT of T4T* was significantly longer than the mean value of RT of T3T* for NE listeners as well. A reasonable explanation for the results for NE listeners in this sense might be that NE listeners relied on multiple acoustic cues such as phonation mode and/or duration in perceiving T3T* and the processing load was thus decreased compared with perceiving T4T* due to the fact that T3 is the only tone that has a creaky phonation and furthermore, that T3 has the longest duration compared to the other three tones (especially to T4) in MC. Obviously, it is not reasonable to count out the possibility that NS listeners used multiple perceptual cues in perceiving T3T* as well[19]. However, there is more reason to assume that a multiple choice in pitch patterns on the second syllable associated with a high initial stem tone triggers an accent 2 processing that involves increased processing load in perceiving T4T* in MC overriding multiple acoustic cues such as phonation mode and/or duration in perceiving T3T* due to the finding that the pre-activation function was detected at 136 ms after stem tone onset (accent 1) for NS listeners (Roll et. al., 2015) that is earlier than the turning point (creaky phonation begins to take place) of T3T* (at 145 ms after the pitch onset, see table 2.1). The result that mean value of RT of T4 as final tone did not differ statistically from any other tones as final tone for NS listeners affords

[19] A few participants from both language groups reported that tone combination T3T1, which was matched with a sheep, sounded actually similar to a sheep, implying phonation mode was used as perceptual cue by them.

further evidence for such an assumption. Additionally, the results showed that mean value of RT of T4T* did not differ statistically from T2T* ($p = .914$) but mean value of RT of T*T4 was significantly shorter than mean value of RT of T*T2 for NE listeners. Thus, for NE listeners, the response time for T2 or T4 as initial tone was similar, but the response time for T4 as final tone was significantly shorter than T2 as final tone, but for NS listeners, the response time for T2 as initial tone was significantly shorter than T4 as initial tone and the difference of response time for T2 or T4 as final tone was not significant statistically. In other words, the processing load for T4 as initial tone for NS listeners and the processing load for T2 as final tone for NE listeners were comparable. The increased processing load for T4 as initial tone for NS listeners can be due to the negative transfer from the increased processing load for accent 2 words in Swedish, the increased processing load for T2 as final tone for NE listeners will be explained in the next section (5.1.1.2).

5.1.1.2 Initial and final tone sub-condition for mismatched trials

There was no effect of initial tone in correct differentiation for NS listeners and there was no effect of L1 background on mean values of neither RA nor RT for mismatched trials. The performance in differentiating was significantly worse than identifying among different tone combinations in both initial and final tone sub-condition for both NS and NE listeners.

NS listeners had only significantly more difficulty to identify whether it was T4 or T3 as initial tone but had significantly more difficulty to discriminate T4 as initial tone from any other tones in the same position while NE listeners had not significantly more difficulty to identify T4 as initial tone from any other tones in the same position but had significantly more difficulty to differentiate T4 as initial tone from T1 and T3 in the same position. The degree of difficulty in differentiating T2 or T4 as initial tone did not differ statistically for NE listeners, but differed statistically for NS listeners. T4 as final tone was not significantly more difficult to identify compared to other tones as final tone but was significantly more difficult

to discriminate than T1 and T3 as final tone for NS listeners while T4 as final tone was significantly more difficult both to identify and to discriminate than T3 as final tone for NE listeners.

Thus, T4 as initial tone was more difficult both to identify and discriminate for NS listeners while T4 as initial tone did not show the same degree of difficulty in perception for NE listeners. A potential explanation of this finding is that the negative transfer of accent 2 words processing in Swedish possibly triggered by high initial stem tone might impede the processing of T4T* and thus lead to comparably lower learning performance in T4T* for NS listeners. For NE listeners, this linguistic experience is missing and thus there was not an effect of T4T*. Nevertheless, the fact that T*T4 was significantly more difficult both to identify and discriminate than T*T3 for NE listeners might be explained by NE listeners relying on multiple perceptual cues to a greater extent compared with NS listeners in identifying and discriminating T*T3. From another perspective, the multiple perceptual cues in perceiving T3T* vs. T4T* was probably overridden by the accent 2 words processing triggered by high initial stem tone for NS listeners. Moreover, the generally lowest perceptual performance of T4T*and T*T4 in terms of solely mean value of response accuracy both in identification and discrimination for both NS and NE listeners with only one exception [T2T*(83.5%) < T4T* (84.7%), for NE listeners] are in accordance with higher perceptual threshold for pitch frequency with shorter duration (Small, 1973:375) and increased JND for pitch frequency with shorter duration in combination with a linear descending ramp in T4 (Klatt, 1973:8) as mentioned earlier in 2.4.1.

In terms of RT, only mean RT of T4T* was significantly longer than mean RT of T1T* and T3T* for NS listeners while there were no significant differences in initial tone sub-condition for NE listeners and in final tone sub-condition for neither NS nor NE listeners. Again, this result is in agreement with Söderström et al. (2012) in that an accent 2 on the stem resulted in

significantly increased response time for verbs with mismatched present tense suffix *-er* in comparison with an accent 1 on the stem resulted in significantly lower response time for verbs with the past tense suffix *-te*. Taken together with the prolonged RT of T4T* for NS listeners in matched trials, the result indicates T4T* was slower processed for NS listeners when sound and image both matched and mismatched each other. Furthermore, results also revealed that the mean value of RT differed non-significantly between T2T* and T4T*($p = .998$) for NE listeners and between T*T4 and T*T2 ($p = .984$) for NS listeners for mismatched trials, in comparison with the mean value of RT of T4T* differed non-significantly from T2T* ($p = .914$) and mean value of RT of T*T4 was significantly shorter than mean value of RT of T*T2 for NE listeners for matched trials. The aforementioned results in terms of mean value of RT in combination with the results that T2T* differed non-significantly from T4T* ($p = 0.981$) in identification and discrimination accuracy for NE listeners, T*T4 differed non-significantly from T*T2 ($p = 0.997$) in identification accuracy for NS listeners, and T*T2 and T*T4 differed non-significantly in discrimination accuracy for either NS ($p = 0.998$) or NE ($p = 1.000$) listeners raise another intriguing aspect of results of this study, namely, the question of confusion in identifying and discriminating between T2 and T4 either as initial or final tone for both NS and NE listeners though the confusion models between identification and discrimination differ to some extent between the two language groups. Krishnan, Gandour, & Bidelman (2010) raised a potentially universal aspect of tone perception in that there may exist a physiological bias toward rising (cf. falling) pitch representation at the brain stem for native speakers of tonal languages, and tonal and nontonal language speakers could be statistically differentiated by the degree of their brain stem response to rising (cf. falling) pitches. Thus, the difficulty to identify between T2 and T4 as initial tone for NE listeners but not for NS listeners might be due to the missing of such a physiological bias toward rising pitch for NE listeners. In consistence with Krishnan et al.

(2010), Burham et al. (2015) found that NS listeners had a physiological bias toward rising pitch comparable to native listeners of a tonal language such as Thai, Cantonese and MC in perceiving five Thai tones. However, the confusion between T2T* and T4T* in terms of mean value of both RA and RT for both NS and NE listeners found in this study is only partly in line with Krishnan et al. (2010) and Burham et al. (2015). It might be speculated that the bias toward rising pitch was masked by the confusion in perceiving T4T* triggered by high initial stem tone in processing of accent 2 words in Swedish for NS listeners.

5.1.2 General learning performance in terms of the easiest and most confusing tone combinations

As reported earlier in 4.3.1.1.1.2, the duplication of the same tone at both initial and final syllable such as T3T3, T1T1, T2T2 were among the easiest tone combinations to learn except for T4T4, which was found among the most confusing tone combinations. Not unexpectedly, decreased memory burden in duplication of the same tone contributes to the ease of learning, but it seems that this easiness of the duplication of the same tone in both initial and final syllable is probably masked by higher perceptual threshold for pitch frequency with shorter duration (Small, 1973:375) and increased JND for pitch frequency with shorter duration in combination with a linear descending ramp (Klatt, 1973:8) in perceiving T4T4. In consistent with So et al. (2010), tone pairs with dissimilar phonetic figures (T1T3 and T3T1) were also found to be easy for both NS and NE listeners in this study. The result that tone pairs such as T3T2 and T2T3 were among the easiest tone combinations to learn found in the present study was not in line with Chen (2012). If the creaky phonation mode and longer duration attributed to the easy differentiation of T3 from all other tones in this study, along with tone pairs involving the rising tone were significantly more discriminable than those involving the falling tone in Burnham et al. (2015), it is not likely that the same reason can explain why T2T3 was the most difficult tone pair to differentiate for NE listeners in Chen (2012) because natural stimuli were used in this study as well as in Chen (2012). Perhaps the association

between the pitch patterns and the images in AI training paradigm adopted in this study facilitates the differentiation between T2 and T3 compared to the traditional auditory training paradigm adopted in Chen (2012).

The T4T*, T1T2, T2T1 and T2T4 were the most confusing tone combinations to learn for both NS and NE listeners, but the order of the difficulty among these tone combinations differed between NS and NE listeners based on the influence of L1 background in prosody as explained in 5.1. Interestingly but not unexpectedly, NS listeners outperformed NE listeners in all tone combinations except T4T2 for all trials and T4T2 and T4T3 for matched trials.

5.2 Effects of AI training paradigm on learning performance

The results indicated that NS listeners gained a substantial increase of 24.2 percent units in overall tone perception accuracy (improved from 64.7% in block 1 to 88.9% in block 6) whereas NE listeners gained a substantial increase of 25.4 percent units in overall tone perception accuracy (improved from 60.3% in block 1 to 85.7% in block 6) for all trials as a result of training. Furthermore, both NS and NE listeners learned to identify tones in MC successfully after training according to Leather's (1990) passing criterion for correct identification of lexical tones in MC (80-85%) in perceptual learning. All sixteen tone combinations were correctly identified at a level higher than 80% in block 3 (14.4 minutes) for NS listeners and in block 4 (19.2 minutes) for NE listeners. Overall, NS listeners gained an increase of 22.9 percent units in correct identification (improved from 72.5% in block 1 to 95.4% in block 6) whereas NE listeners gained an increase of 29.1 percent units in correct identification (improved from 66.2% in block 1 to 95.3% in block 6) in the whole training pass consisting of six blocks. This result of increased percent units in correct identification of the learning tones was higher than results reported from previous traditional auditory tone training studies mentioned earlier in the background part (2.5.1), for instance, Wang et al. (1999), So (2006), Wayland et al. (2008) and Francis et al. (2007). Furthermore, compared to

the above-mentioned traditional auditory tone training studies, this present study with AI training approach was more effective in terms of training time, complexity of stimuli (partly), form of feedback, participants' previous exposure to the training tones, among others. The following paragraphs analyse the differences from the aforementioned perspectives.

Firstly, a percentage gain of 29.1% in correct identification for NE listeners in the present study was higher than a percentage gain of 21% for American NE listeners in Wang et al. (1999), a percentage gain of 23.5% in correct identification for simple feedback group of NE listeners in So (2006), a percentage gain between 12.9% to 14.4% in correct identification for NE listeners in Wayland et al. (2008), and an identification accuracy increase of 16.7 percent units for NE listeners in Francis et al. (2007). Due to So (2006), Wayland et al. (2008) and Francis et al. (2007) were cross-linguistic studies of perceptual learning of lexical tones (MC, Thai and Cantonese) by listeners of NE, NJ (native Japanese) and NMC (native Mandarin Chinese), a comparison between NS listeners in this study and NJ, NMC listeners in the abovementioned studies was thus made. For NS listeners, a gain of 22.9 percent units of correct identification in the present study was higher than a perceptual gain of 22.4% in correct identification for simple feedback group of NJ listeners in So (2006), a perceptual gain between 4.2% to 8.8% in correct identification for NMC listeners in Wayland et al. (2008), and an identification accuracy increase of 9.3 percent units for NMC listeners in Francis et al. (2007).

The finding that NE listeners' identification improved to a greater degree than NS listeners across all blocks in the present study is of interest. This result is line with the greater improvement of correct identification for NE listeners compared to NJ listeners in So (2006), and NMC listeners in Wayland et al. (2008) and Francis et al. (2007). The greater improvement for NE listeners could also be observed in SPSS report in which the f-value of NE listeners ($F (5,155) = 18.55$, $p < .001$) on the effect of block was slightly higher than the f-

value of NS listeners ($F(5,155) = 17.93$, $p < .001$). The identification performance in percentage of correct response for NS listeners was 6.3% higher than NE listeners already in block 1. NS listeners made greater improvement in block 3 and between block 3 and block 6 NS listeners did not make significant improvements in correct identification whereas NE listeners made greater improvement in block 4 and until between block 5 and block 6 NE listeners did not make significant improvements anymore in correct identification. This result suggests a possible ceiling effect among NS listeners rather than relatively greater identification ability among NE listeners as a result of training (identification accuracy 95.4% for NS listeners and 95.3% for NE listeners in block 6). In block 1 when participants needed to guess the match and mismatch between the tone combination and animal image, an initial lower level of identification ability at this stage among NE listeners afforded them a greater space for improvement during the training blocks. On the other hand, NS listeners whose initial identification ability was already relatively high in the first block gained the optimal level of improvement in identification accuracy in block 3 (90.5%) through training. With consideration of the significant improvement from block 4 to block 6 ($p = .001$) for all trials for NS listeners but not for NE listeners, it is not difficult to assume that at later training blocks NS listeners focused more on discrimination of tones instead. In other words, NS listeners reached a possible ceiling effect in identification at earlier blocks compared to NE listeners. At the same time, it is reasonable to argue for that NE listeners benefited more from the AI training paradigm than NS listeners because NE listeners made greater improvements in percent units across the training blocks and in this sense the AI training paradigm was more effective for NE listeners than for NS listeners. However, NE listeners had a significant lower overall learning performance with both identification and discrimination included than NS listeners indicated that the AI training paradigm was more effective for NS listeners in overall perceptual learning of tones in MC. The results from the aforementioned perspectives support

the hypothesis mentioned in 2.6.3. That is, the language group that have lower learning performance (NE listeners) is hypothesized to benefit more from the AI training paradigm in the sense that the AI training paradigm is effective in raising tone-meaning awareness for NE listeners. If the opposite results are found, it means that the AI training paradigm is not effective in raising tone-meaning awareness for native speakers of a non-tone and non-pitch accent language but rather more effective in extending tone-meaning awareness for native speakers of a pitch accent language. The results of this study showed that this was not the case.

Secondly, the present study did not adopt a training paradigm in form of pretest-training-posttest as the above-mentioned traditional auditory tone training studies, but rather employed an on-line training paradigm of learning through guess and simple feedback after only a few minutes' familiarization with the training procedure. It took 14.4 minutes for NS listeners and 19.2 minutes for NE listeners to reach the correct identification level of 80-85% and it took 28.8 minutes for both NS and NE listeners to reach a level of higher than 95% in correct identification in the final block, indicating that the present study with AI training approach was more effective in terms of training time compared to the training period of one hour per day for two days in Wayland et al. (2008), 40 min per day for two weeks in Wang et al. (1999), between 1.25 to 2.5 training hours for simple feedback group to reach a level of 80% correct identification in So (2006), and one hour per day over the course of 10 days for NMC listeners and one hour per day over the course of 16 and 30 days for NE listeners in Francis et al. (2007).

Thirdly, the stimuli of sixteen bisyllabic tone combinations presented counterbalanced across six training blocks in the present study were more complex in terms of numbers of tone patterns every stimulus contained compared to monosyllabic words presented pairwise in a successive order of difficulty of tone contrasts in Wang et al. (1999), five monosyllabic

minimal pairs of low and mid tones of standard Thai presented per trail at first training phrase and in random order at second training phrase in Wayland et al. (2008), twelve monosyllables distributed evenly among six sets of exercises in So (2006), and six target monosyllables each carrying each of the six Cantonese tones presented in random order in Francis et al. (2007). On the other hand, it is worth to note that there was one and the same vowel in two of the similar syllable structure of all sixteen bisyllabic stimuli in the present study while there were different initial consonants and final vowels carrying different tones in monosyllabic stimuli [six tones in Francis et al. (2007) and four tones in So (2006)], and there were various initial consonants and final monophthongs and diphthongs in monosyllabic stimuli in Wang et al. (1999) and Wayland et al. (2008). Thus, the burden of auditory memory was more on the level of suprasegments (pitch patterns) and minimal on the level of segments (*pata* in sixteen tone combinations) in the present study while it was hard to exclude the potential that participants could be distracted by different syllable structures at segment level when perceiving different tones in the other four aforementioned previous studies. Furthermore, all stimuli were produced by a single speaker in the present study while stimuli were produced by different speakers in the other aforementioned studies because one of the motivations of the present study was to investigate the learnability of different pitch patterns for lexical purpose, not the generalization to new untrained stimuli as in the other aforementioned studies. Thus, the participants in the present study avoided being adversely affected by acoustic variations produced by different speakers that were less relevant to the identification of new acoustic categories as in the other aforementioned studies. According to Tone Normalization Theory (Shi, 1986), tone types are relative pitch patterns. Native speakers of a tonal language categorize one tone with slightly different acoustic variations produced by different speakers still as the same tone as long as the contrasts between different tones are categorical, but this categorical perception is hard to reach for adult non-native, naïve listeners.

Fourthly, a simple feedback (lasted in 1.5 s) in form of correct/incorrect response was given after every response in the present study seems to be effective in comparison to the same stimulus was repeated twice first in embedded neutral English voice (e.g. This was Tone 3) and later in isolation after the stimulus (e.g bei 3) was presented and participant's response was given in Wang et al. (1999), the stimulus replay was allowed and the correct response was blinking for a period of 5 seconds once the wrong response was given in Wayland et al. (2008), and there was no upper limit of response time and no feedback in Francis et al. (2007). Indeed, an immediate simple feedback in combination with a large amount of repetitions of the stimuli in quick presentation in this study seems to be effective in establishing the association between the categories varied in pitch patterns and the corresponding lexical meanings and thus contributes in sustaining participants' interest, motivation and engagement generally (Schremm et al., 2016). It is argued for, from one perspective, that immediate responses under time pressure reflect true perceptual process and are comparatively free from post-perceptual process (Cutler, Dahan & Van Donselarr,1997). From another perspective, a quick simple feedback also set higher demands on participants' quick response ability which is not optimal for every participant. Some participants stated that they could perform better if ITI of the present study could be a little longer in their self-report after the experiment. Moreover, it is worth to point out that participants in Wang et al. (1999) were native American English learners of Mandarin Chinese while participants in this study and in other aforementioned studies were naïve listeners in the studying language.

Thus, in sum, despite the training paradigm in the present study adopted shorter training time, more complex stimuli in terms of pitch patterns, only simple feedback and short ITI, listeners' learning performance in correct identification still surpassed the other four aforementioned lexical tone perceptual learning studies that employed traditional auditory tone training paradigm. It is worth mentioning that the result of this study was based on counting

participants' musical aptitude as covariate and no participants received formal musical training prior to or during the time of the study in So (2006), thus the results of these two studies are comparable in this sense. However, participants' musical aptitude was not taken into consideration in neither Wang et al. (1999), nor Wayland et al. (2008), nor Francis et al. (2007). Thus, it is difficult to know how a potential confounding factor such as participants' musical aptitude influenced the results of these three studies. Nevertheless, previous traditional auditory tone training studies showed that non-native listeners can improve in identifying lexical tones in nonlexical contexts. That is, to learn to solely perceive the trajectory of the pitch patterns without putting them in contrast of lexical meaning of words. On the other hand, methodologically, the present study with AI training paradigm showed that native listeners of both a pitch accent language and a non-lexical-tone and non-pitch-accent language can comparably more effectively and more successfully learn to use of pitch patterns for lexical identification.

Wong et al. (2007) employed phonetic–phonological–lexical continuity training paradigm similar to this study and suggested the importance of both increasing non-native listeners' phonological awareness and general acoustic ability in terms of pitch height and pitch pattern in non-native perceptual study of lexical tones (MC). What differed between these two studies methodologically was that a performance-based training termination criterion was adopted in Wong et al. (2007) while a time-based criterion was used in the present study. Furthermore, participants (NE listeners) heard each word for four times with its corresponding picture before they quizzed over the words they just learned through feedback (correction was given after wrong response) and it took 7.22 training sessions (30 minutes per session, three to four sessions per week) for successful learners and 9.38 training sessions for less successful learners to achieve the same identification level of 95% in Wong et al. (2007). Thus, the present study was more effective than Wong et al. (2007) in that it only took 28.8 minutes for

both NS and NE listeners to reach a level higher than 95% in correct identification of tones in MC even though similar training paradigm was adopted by both studies. The high efficacy of learning in this study could probably be explained in two aspects. First, natural stimuli with multiple acoustic cues for tone perception such as pitch, intensity, voice quality and duration included were adopted in the present study compared to synthesized stimuli in which pitch was the sole perception cue in Wong et al. (2007). The results of this study showed that natural stimuli with multiple perception cues such as phonation mode, duration could enhance learning. Second, a simple syllable structure at segment level and the same category of images (animals) gave minimal burden of working memory in the present study whereas different syllable structures and their corresponding pictures of different categories in Wong et al. (2007) might give more burdens to working memory potentially.

Although the nonlexical training part of Wong et al. (2007) was similar to Wang et al. (1999), the result of lower increase of 12.95 percent units of correct identification with lexical training (participants' mean nonlexical pitch identification before training was 78.05%) in Wong et al. (2007) compared to an increase of 21 percent units of correct identification in Wang et al. (1999) was conceivable due to the extra requirements on working memory when associating the pitch pattern to its corresponding lexical meaning in Wong et al. (2007). What makes it arguable was the results of higher increase of 29.1 percent units in correct identification in lexical context in this study compared to an increase of 21 percent units of correct identification in non-lexical context in Wang et al. (1999). One reasonable argument lies in that the association between pitch patterns and animal images in AI training paradigm adopted in this study facilitated the working memory of different tone combinations due to different animal images gave meanings to otherwise for non-native speakers of a tonal language's comprehension "meaningless" pitch patterns and therefore increased their phonological awareness of these pitch patterns and thus they eventually realized the importance of the

corresponding match between pitch pattern and lexical meaning at word level that is totally missing or less wide spread in degree and scope in their native languages. When being asked for the strategies for remembering the sixteen tone combinations in a questionnaire after the experiment, majority of the participants stated the association between pitch contours and images was their main strategy. For instance, the image of a happy dog (pa2ta3) in the experiment lifted his head and his tail was curling and this reminded the pitch contour of tone 2 in combination with tone 3 (˧˥) or the cat (pa2ta2) lifted her front claw and it reminded a rising pitch of tone 2 (˧).[20] Hence, the AI training paradigm of this study is optimal compared to previous traditional auditory tone training studies in that it not only shapes the phonological awareness of different patterns for naïve listeners of a non-pitch accent non-tonal language and extends this phonological awareness for naïve listeners of a pitch accent language, but also facilitates their acoustic ability in identifying different pitch patterns through the association between the concrete trajectory of movements of an image and the abstract trajectory of movements of a pitch pattern .

In addition, as expected, the general results for all trails and matched trials showed that both language groups made significant improvements block-by-block in terms of decreased mean values of RT across all blocks. Basically, mean values of RT and RA across blocks for NS and NE listeners for both all trials and matched trials showed an inverse relationship. Thus, the effectiveness of AI training paradigm in terms of the block-by-block improvements can be evaluated in both the increase of RA and decrease of RT. However, since few data on RT in the aforementioned studies are available, it is not possible to compare their effectiveness with the present study. Based on the results that the mean values of RT for NS listeners were longer than NE listeners for all training blocks with the exception in block 3 in which RT of

[20] The association between the pitch patterns and the images differed among participants.

NE listeners was longer, it is tempting to presume that the perceptual learning performance of NS listeners was influenced by their L1 background in prosody.

5.3 Effects of sound and image match/mismatch on learning performance

The results showed that matched trials (only including identification) scored significantly higher than mismatched trials in terms of mean value of RA and matched trials were faster than mismatched trials in terms of mean value of RT for both language groups, suggesting ID training procedure was easier and thus more effective for naïve listeners. This result supports Lindblad's (2000) statement concerning human discrimination of sounds in that – "in the process of sound perception, we identify and categorize different sounds rather than discriminate the difference of the two sounds" as naïve listeners (mentioned earlier in 2.5.2.).

Despite matched trials scoring significantly higher than mismatched trials in terms of mean value of RA for both NS and NE listeners, the performance of NS listeners was significantly better than NE listeners only in matched trials, not in mismatched trials. Although NS listeners outperformed NE listeners in mismatched trials as well, the difference between the two language groups was not significant. This result implies that there might be a possibility that if the proportion of mismatched trials had increased and matched trials had decreased in the experimental design of this study, NS and NE listeners would not have differed significantly in learning performance of tones in MC. Thus, ID training procedure is indispensable for on-line investigating influence of L1 on non-native perceptual learning of tone combinations of disyllabic words in MC at naïve level.

Interestingly, the results showed that mean value of RA was higher for NS listeners than for NE listeners for both matched and mismatched trials, it is thus logic to presume that mean value of RT was shorter for NS listeners than for NE listeners for both matched and mismatched trials since shorter mean value of RT reflects participants' easiness in responding to the match and mismatch between the tone combination and its corresponding image.

However, the opposite result is found, that is, mean value of RT was longer for NS listeners than for NE listeners for both matched and mismatched trials instead. Just as SLM has predicted (2.6.2), this generally illogic result might imply an interference of L1 background as explained in 5.1.

5.4 Influence of musical aptitude on learning performance

5.4.1 Influence of musical aptitude on improvements of RA and RT for all and matched trials across blocks

Generally, the effect of block, tone combination and L1 background became weaker with musical aptitude as covariate compared to without in terms of mean values of RA for both all trials and matched trials, and the effect of musical aptitude on mean values of RA was greater for NS listeners than NE listeners for both all trials and matched trials. The effect of tone combination on mean values of RA across blocks disappeared for NE listeners for all trials and for both language groups for matched trials with musical aptitude as covariate compared to without, indicating that the effect of tone combination on identification accuracy for both language groups and on the whole learning performance of NE listeners is dependent on the musicality of participants.

In terms of influence of musical aptitude on mean values of RT, there was only effect of musical aptitude for NS listeners (not for NE listeners) in all trials and matched trials across blocks. The effect of block became slightly weaker and the effect of tone combination disappeared for all trials and matched trials and for both language groups with musical aptitude as covariate compared to without, indicating that the effect of tone combination on RT is dependent on the musicality of the participants and the effect of block on RT is dependent on the musicality of the participants on a small scale.

5.4.2 Influence of musical aptitude on initial and final sub-condition

In terms of RA, effect of initial tone for matched trials for NE listeners and the effect of both initial and final tone for mismatched trials for both language groups disappeared, effect of L1

background for matched trials became weaker and only the effect of initial tone for matched trials for NS listeners remained with musical aptitude as covariate compared to without, indicating the effect of initial tone in identification accuracy for NS listeners is the only effect that is not dependent on musicality of participants. The effect of musical aptitude on RA was greater for NS listeners than NE listeners in matched trials and further, there was an effect of musical aptitude and an interaction of initial tone × musical aptitude in mismatched trials for NS listeners but not for NE listeners, indicating that musicality of participants significantly influences the accuracy of both identification and discrimination of tones in MC for NS listeners, but not the discrimination accuracy for NE listeners. The greater influence of musical aptitude for NS listeners compared to NE listeners may partly explain why NS listeners are quicker learners than NE listeners in that participants with higher musical aptitude and longer time of musical training are generally more sensitive to subtle changes in musical pitch and this sensitivity in music pitch enhances language-related pitch processing as shown in this study as well as in Wong et al. (2007) and Bidelman et al. (2010, 2011). Nevertheless, musical aptitude has a smaller influence on NS listeners' the quick learning of tones in MC compared to the contribution of their higher awareness of phonetic–phonological–lexical continuity as reported in 4.1.1.

Similar to musical aptitude on mean values of RA, the effect of initial tone in terms of mean values of RT for matched trials for NS listeners was the only effect that was not dependent on musicality of participants. Musicality of participants influenced the mean values of RT on identification of tones in MC for NS listeners but not on discrimination of tones in MC for neither language groups.

Thus, the influence of L1 background in terms of effect of initial stem tone on both mean values of RA and RT for NS listeners in this study exists regardless the musicality of participants was counted as covariate or not. As mentioned earlier in 2.4.2, this study

predicted that participants with higher musical aptitude would generally have a higher performance in perceiving tones in MC than participants with comparatively lower musical aptitude. Despite this factor being excluded, the effect of initial stem tone on both mean values of RA and RT for NS listeners still remains, indicating that only high initial tone induced onto word stems are specified with accent in Swedish and its pre-activation of the incoming suffixes is intrinsic for its native speakers (Roll, et al., 2010), despite of individuals' musicality. Thus, this study shows that this anticipatory function of hearing the stem tone of the stressed syllable of a word for NS listeners is transferable in non-native perception of lexical tones in MC. This result is in agreement with Wang et al. (2011) in that the initial F0 was transferred as acoustic cue in non-native perception of lexical tones in MC by native Korean learners of MC at beginners' level. Furthermore, the level of dynamics in terms of both pitch and intensity range in MC has a larger overlap with the human hearing range of music making MC a more suitable studying language compared to other tonal language such as Thai due to the higher comparability between the processing of pitch in speech and music in question.

5.5 Other factors of concern

The hearing of participants was based on self-report instead of being tested on-site, participants' auditory working memory was not tested on-site, and the scale of participants' musical aptitude was based on participants' self-reflection in combination with the years of musical training and the onset age of musical training instead of being tested on-site. Therefore, it is not certain whether there is a significant difference on participants' auditory working memory between the two language groups in this study and to what extent the better learning performance of NS listeners compared to NE listeners is influenced by perhaps better auditory working memory besides phonological awareness even though age as covariate had been tested and shown negligibly little influences on the general learning performance and the

effects of L1 background on learning performance. A research design in defining and measuring variables such as musicality is necessary to determine to which extent NS listeners' higher musical aptitude has influenced their better learning performance in comparison to NE listeners at a fine-grained level with the consideration of the overlap of hearing range between human speech and music. Moreover, more studies concerning NS listeners' perceptual learning on other tonal or pitch accent languages need to be carried out to verify the generality of the effect of high initial tone and to exclude its peculiarity in perceptual learning of tones in MC due to the pitch height of the two initial stem tone in Swedish is comparable to T3T* and T4T* in MC. Another potential direction in future research to investigate this issue is to employ and extend Gottfried & Suiter (1997) training paradigm by manipulating the stimuli as intact, onset-only, center-only and silent-center during a time course of a syllable to further pinpoint NS listeners' perceptual cue in perceiving tones in MC.

In addition, only native speakers of Finland Swedish (a variety of Swedish with no pitch accent distinction) were excluded from participating in this study. The including of native speakers of different varieties of Swedish (with relatively early accent timing and distinction between focal and non-focal accentuation such as Svea and Göta dialect zones, and with late accent timing and without distinction between focal and non-focal accentuation such as South, Gotland-Dala and North dialect zones) might constitute a confounding factor that could potentially influence mean values of both RA and RT for NS listeners in this study. It would thus be worth employing only native speakers of Central Swedish (or native speakers of the same dialect zone of Swedish) as participants for further investigating the issue in future research.

6. Conclusion

As mentioned in the introduction, the motivation of this study was twofold: to investigate the influences of learners' L1 prosodic system (theoretically) and to examine the efficacy of an AI training paradigm (methodologically) on the performance of learners' naïve perception of lexical tones. Two research questions were raised in the introduction and now, after the analyses and discussions of the results, a general understanding of them will be given below.

Research question 1: How does language background affect naïve and non-native perception of lexical tones?

The results showed that NS listeners outperformed NE listeners with a higher overall mean value of response accuracy in the behavioural task in this study, supporting the hypothesis that the experience of a pitch accent language (NS listeners) facilitates perceptual learning of lexical tones (MC) compared to experience of nontonal language (NE listeners) at a naïve level. Moreover, the results also showed an effect of initial tone for the identification of tones in MC for NS listeners, but not for NE listeners. The perceptual difficulty in terms of lower response accuracy and longer response time of tone combinations with T4 as initial tone for NS listeners implies a negative transfer of accent 2 words perception triggered by a high initial stem tone in perceiving tone combinations with T4 as initial tone in MC and thus supports the hypothesis that NS listeners could also be restricted in naïve perception of tones in MC due to the interference from a native tone in perceiving a similar but not entirely identical non-native one.

The second research question is: Is auditory-image (AI) training paradigm effective in naïve listeners' perceptual learning of tones in MC?

The improvements in terms of increased mean values of RA and decreased mean values of RT for both language groups across blocks imply the effectiveness of auditory-image (AI)

training paradigm on perceptual learning of lexical tones in MC. Higher substantial increase in percentage units of overall tone perceptual accuracy and shorter training time gained from an auditory-image (AI) training paradigm adopted by this study compared to traditional auditory tone training paradigm without lexical context adopted by previous studies suggest the efficacy of a training paradigm with phonetic–phonological–lexical continuity. The results also showed that NE listeners gained lower overall score but higher substantial increase in percentage units across blocks compared to NS listeners, indicating auditory-image (AI) training paradigm is more effective in raising tone-meaning awareness for NE listeners rather than extending tone-meaning awareness for NS listeners. NS listeners outperformed NE listeners both in matched trials (significantly) and mismatched trials (non-significantly), and both NS and NE listeners passed Leather's criteria of 80-85% correct identification level for successful learning of tones in MC indicate identification training procedure is effective for investigating the efficacy of naïve non-native perceptual learning of MC and the influence of L1 backgrounds in prosody on naïve perceptual learning of MC as well.

As for the influence of musical aptitude on learning performance, the results showed that musical aptitude could be one of the elements to speed up the learning process for identifying tones in MC and it caused the effects of L1 background on overall identification accuracy weaker. The effect of tone combination on identification accuracy for both language groups and on the whole learning performance of NE listeners was dependent on the musicality of participants. However, the effect of initial stem tone on both mean values of RA and RT in identification accuracy for NS listeners in this study was not dependent on the musicality of participants. And thus, this perspective of the findings in this study further supports the assumption that accent 1 and accent 2 in Swedish might be processed differently online. Furthermore, it affords extra evidence to the previous findings in that the processing load for accent 2 words was greater than accent 1 words (Roll et al., 2010,2015,2017; Söderström et

al., 2012). In addition, this perspective of the findings shows that native Swedish speakers' intrinsic use of pitch accent to pre-activate suffixes can be transferred in perceiving pitch patterns in a lexical tone language such as MC at a naïve level (Schremm et al., 2016).

Several interesting questions remain for further research. The confusion between T2 and T4 for both language groups in this study was unexpected. It would thus be interesting to carry out a categorical perceptual study with both a rising falling continuum and a falling rising continuum and to investigate on which psychophysical or linguistic boundaries listeners of different L1 background in prosody rely in their naïve perception of rising and falling tones in MC. Another interesting question for further research is to examine the inattentive cross-linguistic processing of lexical tones by using MMN to elicit the non-native perception of pitch variations even in the absence of attention that is difficult to catch in behavioural studies. Moreover, an experiment to investigate the neural correlates of phonetic–phonological–lexical learning by participants of different L1 background in prosody would also be interesting to conduct.

7. References

Alexander, J.A., Wong, P.C.M., & Bradlow, A.R. (2005). Lexical tone perception by musicians and non-musicians. In *Proceedings of Interspeech 2005*. Lisbon, Portugal.

Bilderman, G. M., Gandour, J., & Krishnan, A. (2010). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brain stem. *Journal of Cognitive Neuroscience, 23 (2),* 425-34.

Bilderman, G. M., Gandour, J., & Krishnan, A. (2011). Musicians and tone language speakers share enhanced brain stem encoding but not perceptual benefits for musical pitch. *Brain and Cognition, 77 (1),* 1-10. doi: 10.1016/j.bandc.2011.07.006.

Best, C. T. (1995). *Speech perception and linguistic experience: Issues in cross-linguistic research.* Timonium, MD: York Press, 171-204.

Braun, B. & Johnson, E. (2011). Question or tone 2? How language experience and linguistic function guide pitch processing. *Journal of Phonetics 39,* 585–594. doi: 10.1016/j.wocn.2011.06.002.

Bruce, G. (1977). *Swedish word accent in sentence perspective*. Lund: Gleerups.

Bruce, G. & Gårding, E. (1978). A Prosodic Typology for Swedish Dialects. In Gårding, E.,Bruce, G., Bannert, R. (eds.), *Nordic Prosody*. Lund: Gleerup, 219-228.

Burnham, D., Elizabeth F., Di Webster, S. L., Francisco L., & Chayada, A. (1996). Facilitation or attenuation in the development of speech mode processing? Tone perception over linguistic contexts. In Paul McCormack & Alison Russell (eds.), *Proceedings of the Sixth Australian International Conference on Speech Science and Technology*, 587–592. Canberra: Australian Speech Science and Technology Association.

Burnham, D. & Brooker, R. (2002). Absolute pitch and lexical tones: Tone perception by non-musician, musician, and absolute pitch non-tonal language speakers. In Hansen J. & Pellom, B. (eds), *The Seventh International Conference on speech science & technology*. Canberra: Australian Speech Science & Technology Association, 86-91.

Burnham, D., Kasisopa, B., Reid, A., Lacerda, F., Attina, V., Rattanasone, N., X., Schwarz, I. & Webster, D. (2015). Universality and language-specific experience in the perception of lexical tone and pitch. *Applied Psycholinguistics, 336,* 1450-1491. doi: 10.1017/S0142716414000496.

Chandrasekaran, B., Krishnan, A., Gandour, J. (2007). Neuroplasticity in the processing of pitch dimensions: a multidimentional scaling analysis of the mismatch negativity. *Restorative Neurology & Neuroscience, 25*(3/4), 195-210.

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

Chang, S. (2012). Effects of Fundamental Frequency and Duration Variation on the Perception of South Kyengsang Korean Tones. *Language and Speech, 56(2),* 211-228. doi: 10.1177/0023830912443951.

Chen, G. T. (1974). The pitch range of English and Chinese speakers. *Journal of Chinese Linguisitics, 2,* 159-171.

Chen, Z. (2007). An Analysis of Chinese Tones Acquisition and Error Analysis of Japanese, Thai and Korean Students. MA thesis. Shanxi Normal University. 陈子悠. (2007). *日泰韩留学生汉语声调习得及偏误分析研究,* 硕士论文, 陕西师范大学.

Chao, Yuan-ren. (1930). A system of tone letters. *Le Maître Phonétique, 45*. 24-47.

Chen, Yu. (2012). An experimental study of perceiving Mandarin tones for English speakers without Chinese learning experience. *Journal of Hechi University*, 2012, No 3.

Cutler, A., Dahan, D. & Van Donselarr, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language & Speech, 40 (2),* 141-201.

Denes, P. B. & Pinson, E. N. (1993). *The speech chain: the physics and biology of spoken language.* New York: Freeman and Company.

Dong, Y., Tsubota, Y. & Dantsuji, M. (2013). Difficulties in perception and pronunciation of Mandarin Chinese disyllabic word tone acquisition: A study of some Japanese University students. In: Proceedings of the 27th Pacific Asia Conference on language, information, and computation, p. 143-152.

Duanmu, San. (2007). *The phonology of standard Chinese*. Oxford: Oxford University Press.

Elert, Claes-Christian. (1971). *Tonality in Swedish: Rules and a list of minimal pairs*. Umeå: Umeå University Department of Phonetics.

Felder, V., Jönsson-Steiner, E., Eulitz, C., & Lahiri, A. (2009). Asymmetric processing of lexical tone contrast in Swedish. *Attention, Perception & Psychophysics, 71(8),* 1890-1899. doi: 10.3758/APP.71.8.1890.

Flege, J.E. (1995). Speech perception and linguistic experience: Issues in cross-linguistic research. Timonium, MD: York Press, 233-277.

Fromkin, V.A. (ed.) (1978). Tone: A Linguistic Survey. London: Academic P., cop.

Francis, A.L., Ciocca, V., Ma, L. & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and none-tone language speakers. *Journal of Phonetics 36,* 268-294. doi: 10.1016/j.wocn.2007.06.005.

Fu, Q., Zeng, F., Shannon, R.V. & Soli, S.D. (1998). Importance of tonal envelope cues in Chinese speech recognition. *Journal of the Acoustical Society of America, 104 (1),* 505-510. doi.org/10.1121/1.423251.

Gao, Man. (2016). Perception of lexical tones by Swedish learners of Mandarin. *In Linköping Electronic Conference Proceedings 130*: 33-40.

Gandour, J., Harsman, R. (1977). Cross-language difference in tone perception: a multidimentional scaling investigation. *Language Speech, 21,* 1-33.

Gandour, J. (1978). The perception of tone. In Fromkin V.A. (ed). (1978). *Tone: A linguistic Survey.* New York: Academic Press, 41-76.

Gandour, J. T. (1983). Tone perception in far eastern languages. *Journal of Phonetics 11, 149-175.*

Gandour,J. (1984). Tone discrimilarity judgments by Chinese listeners. *Journal of Chinese Linguistics, 12,* 235-261.

Gottfried, T.L. & Suiter, T.L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of phonetics, 25 (2),* 207-231.

Gravetter, F. J., & Forzano, L. A. B. (2018). *Research methods for the behavioral sciences (4th ed.).* Stamford, CT: Cengage Learning.

Guo, R., Wang, & Lu, J. (2012). Modern Chinese. Beijing: Commercial Press. 郭锐，王理嘉, 陆俭明. (2012). *《现代汉语》*, 北京: 商务印书馆.

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

Gårding, E. (1986). Tone 4 and tone 3 discrimination in Modern Standard Chinese. *Language and Speech, 29 (3),* 281-293.

Hadding, k. & Petersson, L. (1970). *Experimentell fonetik.* Lund: Gleerups.

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

Ingram, J (2007). *Neurolinguistics. An introduction to spoken language processing and its disorders.* Cambridge: Cambridge University Press.

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrasts by feancophones. *Perception and Psychophysics, 40 (4),*205-215.

Kaan, E., Barkley, C.M., Bao, M., Wayland, R. (2008). Thai lexical tone perception in native speakers of Thai, English and Mandarin Chinese: An event-related potentials training study. *BMC Neuroscience. 15 (4),* 335-354. doi: 10.1186/1471-2202-9-53.

Kiriloff, C. (1969). On the auditory discrimination of tones in Mandarin. *Phonetica, 20,* 63-67.

Klatt, D (1973). Discrimination of fundamental frequency contours in synthetic speech duplication for models of pitch perception. *Journal of the Acoustical Society of America, 53,* 8-16.

Kong, J. & Zhang, R. (2017). VAT of the lexical tones in Mandarin Chinese. *The Journal of Chinese Linguistics, 45,* 275-289.

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

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

Ladefoged, P & Johnson, K. (2011). *A course in phonetics.* Boston: Wadsworth.

Ladefoged, P. (2003). *Phonetic data analysis.* Malden: Blackwell Publishing.

Leather J. (1990). *Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers.* In New Sounds 90: Proceedings of the Amsterdam Symposium on the acquisition of second language speech. Edited by J. Leather and A. James (University of Amsterdam).

Lee, L., & Nusbaum, H.C. (1993). Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese. *Attention, Perception, and Psychophysics 53,* 157–165. doi: 10.1080/23273798.2016.1190850.

Lehiste, I. (1970). *Suprasegmentals.* Cambridge: MIT Press.

Li, B., Shao, J. & Bao, M. (2017). Effects of Phonetic Similarity in the Identification of Mandarin tones. *Journal of Psycholinguistic Research, 46,* 107-124. doi: 10.1007/s10936-016-9422-6.

Li, B., & Shuai, L. (2011). Effects of native language on perception of level and falling tones. In *Proceeding of 17th international congress of phonetic sciences.* Hong Kong. 1202-1205.

Li, B. & Zhang, C. (2010). Effects of F0 dimensions in perception of Mandarin tones. *In 7th International Symposium on Chinese Spoken Language Processing (ISCSLP).* 322-325. doi: 10.1109/ISCSLP.2010.5684878.

Li, J & Xu, Y. (2018). Whispered Mandarin has no production-enhanced cues for tone and intonation. *Lingua*. (In press) doi: 10.1016/j.lingua.2018.01.004.

Li, Y. (2016). English and Thai Speakers' Perception of Mandarin Tones. *English Language Teaching, 9,* 122-132.

Liang, Z.A. (1963). The auditory perception of Mandarin tones. *Acta Physiologica Sinica, 26,* 85-91.

Lin T. & Wang, L. (1992). *A course in phonetics*. Beijing: Peking University Press. 林涛, 王理嘉. (1986). *《语音学教程》*. 北京: 北京大学出版社.

Lin T. & Wang, W. (1985). Tone perception. *Journal of Chinese Linguistics, 2,* 59-69.

Lindblad, P. (2000). *Psykoakustik*. Kompendium. Institutionen för lingvistik, Lunds Universitet.

Liu, S., & Samuel, A.G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. Language and Speech, 47 (2), 109-138.

Lively, S. E., Yamada, R. A., Tokhura, Y. & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America, 96,* 2076-2087.

Lu, S., Wayland, & R., Kaan, E. (2015). Effects of production training and perception training on lexical tone perception – A behavioural and ERP study. *Brain Research, 1624,* 28-44. doi: 10.1016/j.brainres.2015.07.014.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25,* 71-102.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognition of Psychology, 18,* 1–86. doi: 10.1016/0010-0285(86)90015-0.

Milliken, S. (1989). "Why there is no tone sandhi rule in Standard Mandarin", paper presented at the Tianjin International Conference on Phonetics and Phonology, Tianjin Normal University, June 7-10.

Narahara, T. (1985). The Accentual System of the Korean Kyengsang Dialect. In Kuno, S., Lee, I-H, Whitman, J. and Kang, Y-S. (eds), *Harvard Studies in Korean Linguistics*. Cambridge, MA: Department of Linguistics, Harvard University.

Pike, K.L. 1948. *Tone languages.* AnnArbor: University of Michigan Press.

Riad, T., 1998. The origin of Scandinavian tone accents. Diachronica 15, 63-98.

Riad, T. (2009). The morphological status of accent 2 in North Germanic simplex forms. In Vainio, M., Aulanko, R. & Aaltonen, O. (Eds.), *Nordic Prosody X*. Frankfurt am Main: Peter Lang. 205-216.

Riad, Tomas. (2014). *The phonology of Swedish*. Oxford: Oxford University Press.

Rischel, J., 1963. Morphemic tone and word tone in Eastern Norwegian. Phonetica 10, 154-164.

Roll, M., Horne, M., Lindgren, M. (2010). Word accents and morphology – ERPs of Swedish word processing. *Brain Research, 1330,* 114-123. doi: 10.1016/j.brainres.2010.03.020.

Roll, M., Horne, M., & Lindgren, M. (2011). Activating without inhibiting: Left-edge boundary tones and syntactic processing. *Journal of Cognitive Neuroscience, 23 (5),* 1170-1179. doi: 10.1162/jocn.2010.21430.

Roll, M., Söderström, P. & Horne, M. (2013). Word-stem tones cue suffixes in the brain. *Brain Research, 1520,* 116-120. https://doi.org/10.1016/j.brainres.2013.05.013.

Roll, M., Söderström, P., Mannfolk, P., Shtyrov, Y., Johansson, M., Westen, D., Horne M. (2015). Word tones cueing morphosyntactic structure: Neuroanatomical substrates and activation time-course assessed by EEG and fMRI. *Brain & Language, 150,* 14-21. doi: 10.1016/j.bandl.2015.07.009.

Roll, M., Söderström, P., Frid, J., Mannfolk, P., Horne M. (2017). Forehearing words: Pre-activation of word endings at word onset. *Neuroscience letters, 658,* 57-61. doi: 10.1016/j.neulet.2017.08.030.

Schaefer, V. & Darcy, I. (2014). Lexical function of pitch in the first language shapes cross-linguistic perception of Thai tones. *Laboratory Phonology, 5,* 489-522. doi: 10.1515/lp-2014-0016.

Shen, J. (2016). Processing of pitch height information in Mandarin tone perception. PhD dissertation. University of California.

Shen, X. S. (1989). Toward a register approach in teaching Mandarin tones. J. Chinese Language Teachers Association, 24, 27-47.

Shen, X. S. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech, 34(2),* 145-156.

Shibata, T., & Shibata, R. (1990). How much can accent distinguish homophones? Cases of Japanese, English and Chinese. *Mathematical Linguistics, 17,* 317-23.

Schremm, A., Söderström, P., Horne, M. & Roll, M. (2016). Implicit acquisition of tone-suffix connections in L2 learners of Swedish. The Mental Lexicon 11:1, 55-75. doi: 10.1075/ml.11.1.03sch.

Shi, F. (1986). Chinese tone ratio theory. Retrived from: www.phonetics.org.cn/FUploadFile/.../2837018 论语音格局.doc. 2015-03-09. 石峰, 1986,*《论语音格局》*.

Small, A. 1973. Psychoacoustics. In Minifie, M., Hixon, T. & Williams, F (eds.), *Normal aspects of speech, hearing, and language*. New Jersey: Englewood Cliffs, 343-420.

So, C. K. (2006). The influence of L1 prosodic background on the learning of Mandarin tones: patterns of tonal confusion by Cantonese and Japanese naïve listeners. *In Proceedings of the 2005 Canadian Linguistic Association Annual Conference*.

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.

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

Sonesson, B. (1968). The functional anatomy of the speech organs. In Malmberg, B. (ed.), *Manual of phonetics*. Amsterdam: North-Holland, 45-75.

Söderström, P., Roll, M, & Horne, M. (2012). Processing morphologically conditioned word accents. *The Mental Lexicon 7:1,* 77-89. doi: 10.1075/ml.7.1.04soe.

Söderström, P., Horne, M., Frid, J. & Roll, M. (2016). Pre-Activation Negativity (PrAN) in Brain Potentials to Unfolding Words. *Frontiers in Human Neuroscience. 10.* doi: 10.3389/fnhum.2016.00512.

Tamaoka, K., Saito, N., Kiyama, S., Timmer, K. & Verdonschot, R. G. (2014). Is pitch accent necessary for comprehension by native Japanese speakers? – An ERP investigation. *Journal of Neurolinguistics, 27,* 31-40. doi: 10.1016/j.jneuroling.2013.08.001.

Tang, Z.F. (2013). *A course in prosody*. Beijing: Peking University Press. 唐作藩. 2013.*《音韵学教程》*. 北京: 北京大学出版社.

Tsujimura, Natsuko. 2006. *An introduction to Japanese linguistics*. Malden, MA: Blackwell Publishing.

Tsukada, K., Kondo, M., & Sunaoka K. (2016). The perception of Mandarin lexical tones by native Japanese adult listeners with and without Mandarin learning experience. *Journal of Second Language Pronunciation 2:2,* 225–252. doi: 10.1075/jslp.2.2.05tsu.

Whalen, D.H., & Xu Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. Phonetica ,49, 25-47.

Wang, Y., Spence, M.M., Jongman, A. & Sereno, J. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustic Society of America 106 (6),* 3649-3658.

Wang, Y., Jongman, A. & Sereno J. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustic Society of America 113 (2),* 1033-1044.

Wang, Y., Behne, D.M., Jongman, A., & Sereno, J. (2004). The role of linguistic experience in the hemispheric processing of lexical tone. *Applied Psycholinguistics, 25,* 449-466.

Wang X. C. (2013). Perception of Mandarin tones: The effect of L1 background and training. *Modern Language Journal, 97* (1), 144-160.

Wang, Y & Li, J (2011). The Perceptions of Tone 2 and Tone 3 in Mandarin by Korean Native Speakers. *Language teaching and linguistic studies. 2011, 1.* 王蕴佳, 李美京. (2011). 韩语母语者对普通话阳平和上声的知觉. *《语言教学与研究》*, 第 1 期.

Wayland, R.P. & Guion, S. G. (2004). Training English and Chinese Listeners to Perceive Thai Tones: A Preliminary Report. *Language learning 54:4,* 681-712.

Wayland, R. P. & Li. B. (2008). Effects of two training procedures in cross-language of perception of tones. *Journal of Phonetics, 36 (2),* 250-267. doi: 10.1016/j.wocn.2007.06.004.

Weireich, U. (1953). Language in contact. New York: Linguistic Circle of New York.

White, C. M., (1981). Tonal perception errors and interference from English intonation. *Journal of Chinese Language Teachers Association, 16*, 27-56.

White, C.M. (1981). Mandarin tone and English intonation: a contrastive analysis. Unpublished MA thesis. The University of Arizona.

Wong, P.C.M. & Perrachione, T.K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics 28,* 565–585. doi: 10.1017/S0142716407070312.

Xu, Y., Gandour, J., Talavage, T., Wong., D., Dzemidzic, M., Tong, Y., et al. (2006). Activation of the left planum temporale in pitch processing is shaped by language experience. *Human Brain Mapping, 27,* 173-183.

Yang, X. (2011). The phonation factor in the categorical perception of Mandarin tones. *Regular Session.* 2204-2207.

Yip, Moria. (2002). *Tone*. Cambridge: Cambridge University.

Zhu, W., Wei, Y., Wu, L. & Wang, J. (2016). The effect of pitch range and tone duration on Chinese tone perception by L2 learners. *Chinese Language learning*, 2016, No 2.

Zsiga, E. & Nitisaroj, R. (2007). Tone features, tone perception and peak alignment in Thai. *Language & speech, 50 (3),* 343-383.

Zee, E. & Greenberg, S. (1977). On the perception of contour tones. *The Journal of the Acoustical Society of America, 62,* 47. (Abstract).

8.Appendix

8.1 Appendix 1

Background information and after-experiment questions

Background information:

1. Name:
2. Gender:
3. Age:
4. Education discipline:
5. Mother tongue (including the variety of the language):
6. Other languages studied and the time of studying:

7. Do you have any previous exposure to any tonal languages (e.g. Chinese, Thai)? Do you have any previous exposure to any pitch accent languages (e.g. Swedish, Japanese)?

After experiment questions:

1. Do you feel the experiment is difficult? Please use 1-5 grade to mark the scale of difficulty (1 = easy, 3 = medium, 5 = hard).

2. Which animals are easy? Which are difficult?

3. How musical talented are you? Have you had any musical training before? For how long? At what age did you begin your musical training?

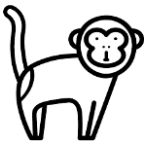
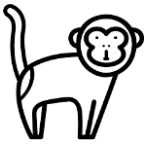
a. Very talented
b. Higher than the middle level
c. Middle level
d. Lower than the middle level
e. Not musical at all

4. How have you tried to remember the animals? Do you imitate the tone combinations during the experiment?

5. How do you pay attention to the tones? Are you more sensitive to pitch onset and offset (a high tone or a low tone at the beginning and the endpoint of the tone) or to the contour differences inside the tone (a rising tone, a falling tone, a first rising then falling tone, ect.)?

8.2 Appendix 2

<table>
<tr><td colspan="9">The design lists with both the matches and mismatches between the sound stimuli and the image stimuli in the experiment (The images were fetched from Google image search/animals)</td></tr>
<tr><td>Images</td><td colspan="2">Duck
  </td><td colspan="2">Chicken
  </td><td colspan="2">Frog
  </td><td colspan="2">Tiger
  </td></tr>
<tr><td>Correct tone combinations</td><td colspan="2">pa1ta1
  </td><td colspan="2">pa1ta2
  </td><td colspan="2">pa1ta3
  </td><td colspan="2">pa1ta4
  </td></tr>
<tr><td rowspan="3">Wrong tone combinations</td><td>pa1ta2
  </td><td>pa2ta1
  </td><td>pa1ta1
  </td><td>pa2ta2
  </td><td>pa1ta1
  </td><td>pa2ta3
  </td><td>pa1ta1
  </td><td>pa2ta4
  </td></tr>
<tr><td>pa1ta3
  </td><td>pa3ta1
  </td><td>pa1ta3
  </td><td>pa3ta2
  </td><td>pa1ta2
  </td><td>pa3ta3
  </td><td>pa1ta2
  </td><td>pa3ta4
  </td></tr>
<tr><td>pa1ta4
  </td><td>pa4ta1
  </td><td>pa1ta4
  </td><td>pa4ta2
  </td><td>pa1ta4
  </td><td>pa4ta3
  </td><td>pa1ta3
  </td><td>pa4ta4
  </td></tr>
<tr><td>Images</td><td colspan="2">Elephant
  </td><td colspan="2">Cat
  </td><td colspan="2">Dog
  </td><td colspan="2">Pig
  </td></tr>
<tr><td>Correct tone combinations</td><td colspan="2">pa2ta1
  </td><td colspan="2">pa2ta2
  </td><td colspan="2">pa2ta3
  </td><td colspan="2">pa2ta4
  </td></tr>
<tr><td>Wrong tone combinations</td><td>pa2ta2</td><td>pa1ta1</td><td>pa2ta1</td><td>pa1ta2</td><td>pa2ta1</td><td>pa1ta3</td><td>pa2ta1</td><td>pa1ta4</td></tr>
</table>

								
	pa2ta3 	pa3ta1 	pa2ta3 	pa3ta2 	pa2ta2 	pa3ta3 	pa2ta2 	pa3ta4 
	pa2ta4 	pa4ta1 	pa2ta4 	pa4ta2 	pa2ta4 	pa4ta3 	pa2ta3 	pa4ta4 
Images	Sheep 		Panda 		Lion 		Monkey 	
Correct tone combinations	pa3ta1 		pa3ta2 		pa3ta3 		pa3ta4 	
Wrong tone combinations	pa3ta2 	pa1ta1 	pa3ta1 	pa1ta2 	pa3ta1 	pa1ta3 	pa3ta1 	pa1ta4 
	pa3ta3	pa2ta1	pa3ta3	pa2ta2	pa3ta2	pa2ta3	pa3ta2	pa2ta4
	pa3ta4	pa4ta1	pa3ta4	pa4ta2	pa3ta4	pa4ta3	pa3ta3	pa4ta4
Images	Fish		Bear		Rabbit		Cow	

Correct tone combinations	pa4ta1		pa4ta2		pa4ta3		pa4ta4	
								
Wrong tone combinations	pa4ta2	pa1ta1	pa4ta1	pa1ta2	pa4ta1	pa1ta3	pa1ta4	pa4ta1
								
	pa4ta3	pa2ta1	pa4ta3	pa2ta2	pa4ta2	pa2ta3	pa2ta4	pa4ta2
								
	pa4ta4	pa3ta1	pa4ta4	pa3ta2	pa4ta4	pa3ta3	pa3ta4	pa4ta3
								

8.3 Appendix 3

Table 1 Mean values of RA of each tone combination for all trials across blocks for both language groups (with 95% confidence interval).

		Blocks					
Tone combinations	Language Groups	1	2	3	4	5	6
T1T1	NS listeners	.729	.818	.897	.919	.931	.926
	NE listeners	.622	.758	.804	.870	.871	.889
T1T2	NS listeners	.570	.658	.884	.774	.893	.888
	NE listeners	.558	.743	.779	.811	.788	.826
T1T3	NS listeners	.642	.734	.829	.866	.839	.916
	NE listeners	.610	.691	.812	.818	.874	.908
T1T4	NS listeners	.548	.701	.890	.872	.849	.891
	NE listeners	.595	.643	.738	.779	.850	.855
T2T1	NS listeners	.646	.742	.814	.804	.886	.901
	NE listeners	.623	.638	.711	.806	.878	.833
T2T2	NS listeners	.676	.773	.852	.899	.902	.896
	NE listeners	.594	.707	.793	.847	.891	.855
T2T3	NS listeners	.624	.811	.831	.865	.920	.867
	NE listeners	.613	.705	.800	.778	.855	.857
T2T4	NS listeners	.572	.725	.782	.811	.795	.885
	NE listeners	.589	.594	.658	.765	.810	.834
T3T1	NS listeners	.640	.695	.839	.896	.879	.927
	NE listeners	.600	.736	.848	.796	.821	.831
T3T2	NS listeners	.663	.745	.875	.873	.909	.881
	NE listeners	.576	.743	.736	.887	.860	.822
T3T3	NS listeners	.741	.873	.939	.971	.979	.970
	NE listeners	.651	.765	.876	.930	.959	.905
T3T4	NS listeners	.747	.750	.863	.799	.873	.930
	NE listeners	.591	.754	.774	.812	.866	.873
T4T1	NS listeners	.643	.755	.775	.815	.805	.832
	NE listeners	.507	.717	.799	.839	.795	.839
T4T2	NS listeners	.600	.754	.710	.728	.840	.837
	NE listeners	.613	.680	.690	.839	.850	.825
T4T3	NS listeners	.692	.743	.739	.822	.853	.860
	NE listeners	.630	.696	.697	.844	.828	.822
T4T4	NS listeners	.639	.719	.816	.804	.788	.882

NE listeners	.652	.653	.727	.737	.782	.870

Table 2 Improvements in percentage of correct response between blocks for all trials for each tone combination and language group.

Improvement in percentage of correct response						
Tone combinations	Language Groups	block 1→ 2	block 2→ 3	Block 3→ 4	block 4→ 5	Block 5→ 6
T1T1	NS listeners	9	8	2	1	-1
	NE listeners	14	5	7	0	2
T1T2	NS listeners	9	23	-11	12	-1
	NE listeners	19	4	3	-2	4
T1T3	NS listeners	9	10	4	-3	8
	NE listeners	8	12	1	6	3
T1T4	NS listeners	15	19	-2	-2	4
	NE listeners	5	10	4	7	1
T2T1	NS listeners	10	7	-1	8	2
	NE listeners	2	7	10	7	-5
T2T2	NS listeners	10	8	5	0	-1
	NE listeners	11	9	5	4	-4
T2T3	NS listeners	19	2	3	6	-5
	NE listeners	9	10	-2	8	0
T2T4	NS listeners	15	6	3	-2	9
	NE listeners	1	6	11	5	2
T3T1	NS listeners	5	14	6	-2	5
	NE listeners	14	11	-5	2	1
T3T2	NS listeners	8	13	0	4	-3
	NE listeners	17	-1	15	-3	-4
T3T3	NS listeners	13	7	3	1	-1
	NE listeners	11	11	5	3	-5
T3T4	NS listeners	0	11	-6	7	6
	NE listeners	16	2	4	5	1
T4T1	NS listeners	11	2	4	-1	3
	NE listeners	21	8	4	-4	4
T4T2	NS listeners	15	-4	2	11	0
	NE listeners	7	1	15	1	-3

T4T3	NS listeners	5	0	8	3	1
	NE listeners	7	0	15	-2	-1
T4T4	NS listeners	8	10	-1	-2	9
	NE listeners	0	7	1	5	9

8.4 Appendix 4

Table 1 Mean values of RA of each tone combination for matched trials across blocks for both language groups (with 95% confidence interval).

		Blocks					
Tone combinations	Language Groups	1	2	3	4	5	6
T1T1	NS listeners	.747	.859	.909	.970	.960	.949
	NE listeners	.677	.778	.848	.919	.939	.950
T1T2	NS listeners	.646	.778	.950	.869	.990	.990
	NE listeners	.616	.838	.818	.869	.899	.939
T1T3	NS listeners	.737	.828	.899	.990	.949	.970
	NE listeners	.626	.828	.859	.859	.939	1.000
T1T4	NS listeners	.667	.798	.949	.929	.949	.960
	NE listeners	.646	.707	.848	.848	.939	.919
T2T1	NS listeners	.707	.828	.909	.899	.939	.960
	NE listeners	.667	.687	.808	.809	.960	.919
T2T2	NS listeners	.727	.859	.949	.980	.949	.949
	NE listeners	.646	.798	.909	.919	.960	.950
T2T3	NS listeners	.707	.889	.909	.919	.970	.949
	NE listeners	.657	.778	.869	.848	.949	.960
T2T4	NS listeners	.687	.788	.879	.919	.919	.960
	NE listeners	.626	.697	.788	.889	.909	.939
T3T1	NS listeners	.737	.838	.909	.939	.929	.970
	NE listeners	.717	.798	.929	.899	.929	.949
T3T2	NS listeners	.768	.848	.949	.970	.980	.970
	NE listeners	.636	.808	.838	.949	.909	.939
T3T3	NS listeners	.838	.909	.970	.980	1.000	.990
	NE listeners	.758	.838	.929	.970	.990	.990
T3T4	NS listeners	.828	.899	.960	.980	.919	.980
	NE listeners	.687	.879	.838	.909	.949	.970
T4T1	NS listeners	.677	.848	.859	.919	.949	.929
	NE listeners	.515	.808	.889	.949	.949	.939
T4T2	NS listeners	.667	.828	.848	.869	.939	.909
	NE listeners	.687	.778	.838	.929	.939	.929
T4T3	NS listeners	.758	.808	.808	.899	.939	.909
	NE listeners	.707	.828	.798	.949	.879	.939
T4T4	NS listeners	.717	.778	.889	.899	.889	.960

	NE listeners	.707	.727	.838	.848	.879	.970

Table 2 Improvements in percentage of correct response between blocks for matched trials for each tone combination and language group.

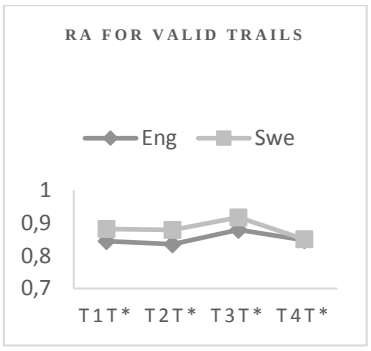
Improvement in percentage of correct response						
Tone combinations	Language Groups	block 1→ 2	block 2→ 3	block 3→ 4	block 4→ 5	Block 5→ 6
T1T1	NS listeners	11	5	6	-1	-1
	NE listeners	10	7	7	2	2
T1T2	NS listeners	13	17	-8	12	12
	NE listeners	22	-2	5	3	3
T1T3	NS listeners	9	7	9	-4	-4
	NE listeners	20	3	0	8	8
T1T4	NS listeners	13	15	-2	2	2
	NE listeners	6	14	0	9	9
T2T1	NS listeners	12	8	-1	4	4
	NE listeners	2	12	0	15	15
T2T2	NS listeners	13	9	3	-3	-3
	NE listeners	15	11	1	4	4
T2T3	NS listeners	18	2	1	5	5
	NE listeners	12	9	-2	10	10
T2T4	NS listeners	10	9	4	0	0
	NE listeners	7	9	10	2	2
T3T1	NS listeners	10	7	3	-1	-1
	NE listeners	8	13	-3	3	3
T3T2	NS listeners	8	10	2	1	1
	NE listeners	17	3	11	-4	-4
T3T3	NS listeners	7	6	1	2	2
	NE listeners	8	9	4	2	2
T3T4	NS listeners	7	6	2	-6	-6
	NE listeners	19	-4	7	4	4
T4T1	NS listeners	17	1	6	3	3
	NE listeners	29	8	6	0	0
T4T2	NS listeners	16	2	2	7	7
	NE listeners	9	6	9	1	1

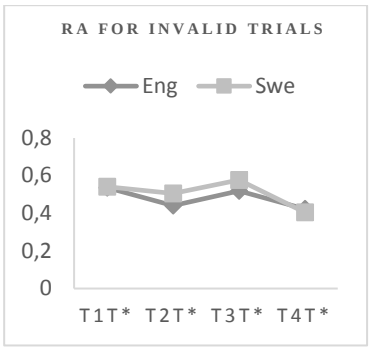
T4T3	NS listeners	5	0	9	4	4
	NE listeners	12	-3	15	-7	-7
T4T4	NS listeners	6	11	1	-1	-1
	NE listeners	2	11	1	3	3

8.5 Appendix 5

Figure 1 Initial and final tone sub-condition: Mean values of RA and RT for all, matched and mismatched trials for both NS and NE listeners.

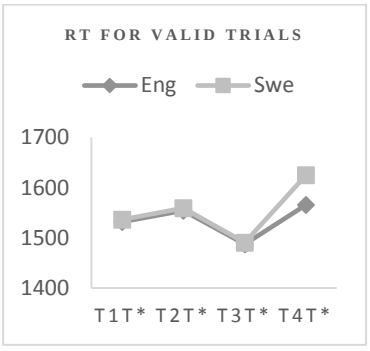
Initial tone sub-condition (RA for matched and mismatched trials)

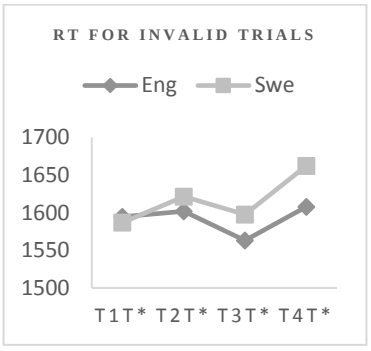




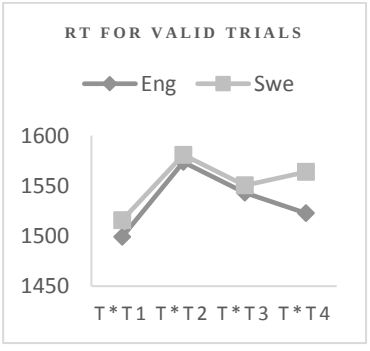
Final tone sub-condition (RA for matched and mismatched trials)

Initial tone sub-condition (RT for matched and mismatched trials)





Final tone sub-condition (RT for matched and mismatched trials)



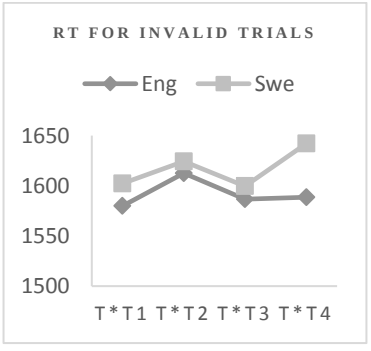


Table 1 Effects of pairwise comparisons of RA of initial tone and final tone sub-condition for all, matched and mismatched trials for NS listeners (upper right triangle, read in vertical axis in terms of initial or final tone sub-condition) and NE listeners (lower left triangle, read in horizontal axis in terms of initial or final tone sub-condition). Confidence interval for difference was 95% and covariate appearing in the model were evaluated at the following values: Musical aptitude = 2.48).

Initial tone sub-condition

	matched trials				mismatched trials			
	T1T*	T2T*	T3T*	T4T*	T1T*	T2T*	T3T*	T4T*
T1T*	—	Ns	***	Ns	—	**\Ns	Ns	*****
T2T*	Ns	—	**	Ns	**\Ns	—	**	Ns***
T3T*	***	**	—	Ns***	Ns	**	—	*****
T4T*	Ns	Ns	Ns***	—	*****	Ns***	*****	—

Final tone sub-condition

	matched trials				mismatched trials			
	T*T1	T*T2	T*T2	T*T4	T*T1	T*T2	T*T2	T*T4
T*T1	—	Ns	Ns	Ns	—	Ns	Ns	Ns**
T*T2	Ns	—	Ns	Ns	Ns	—	*****	Ns
T*T3	Ns	Ns	—	**\Ns	Ns	*****	—	*****
T*T4	Ns	Ns	**\Ns	—	Ns**	Ns	*****	—

Ns: not significant, *p < .05, **p < .01, ***p < .001 (Adjustment for multiple comparisons: Sidak).

Table 2 Effects of pairwise comparisons of RT of initial tone and final tone sub-condition for all, matched and mismatched trials for NS listeners (upper right triangle, read in vertical axis in terms of initial or final tone sub-condition) and NE listeners (lower left triangle, read in horizontal axis in terms of initial or final tone sub-condition). Confidence interval for difference was 95% and covariate appearing in the model were evaluated at the following values: Musical aptitude = 2.48).

Initial tone sub-condition

	matched trials				mismatched trials			
	T1T*	T2T*	T3T*	T4T*	T1T*	T2T*	T3T*	T4T*
T1T*	—	Ns	***	Ns***	—	Ns	Ns	Ns**
T2T*	Ns	—	*****	Ns***	Ns	—	Ns	Ns
T3T*	***	*****	—	***	Ns	Ns	—	Ns*
T4T*	Ns***	Ns***	***	—	Ns**	Ns	Ns*	—

Final tone sub-condition

	matched trials				mismatched trials			
	T*T1	T*T2	T*T3	T*T4	T*T1	T*T2	T*T3	T*T4
T*T1	—	*****	*	Ns	—	Ns	Ns	Ns
T*T2	*****	—	Ns	***\Ns	Ns	—	Ns	Ns
T*T3	*	Ns	—	Ns	Ns	Ns	—	Ns
T*T4	Ns	***\Ns	Ns	—	Ns	Ns	Ns	—

Ns: not significant, *p < .05, **p < .01, ***p < .001 (Adjustment for multiple comparisons: Sidak).