Investigating Spoken Emotion:

The Interplay of Language and Facial Expression

Chong Chee Seng

BPsySc (Hons)

A Thesis Submitted for the Degree of

Doctor of Philosophy

The MARCS Institute

Western Sydney University

March 2019

Acknowledgements

Firstly, I would like to thank my supervisors Jeesun Kim and Chris Davis for the guidance, wisdom and care that you have given me over the years. I am deeply indebted to the two of you for the encouragement and opportunity to start this thesis, which has given me a direction and goal in life.

I am very grateful to the MARCS Institute for providing me with the opportunity to pursue my interests and to the technical support team: Colin Schoknect, Steven Fazio, Lei Jing, Johnson Chen, Donovan Govan and Ben Binyamin for your assistance. I would like to thank the Sunway University's Department of Psychology for giving me the opportunity to visit and to conduct an experiment at your lab. I would also like to thank Vincent Aubanel, Gregory Zelic and Yatin Mahajan. Your advice, knowledge and experience have been invaluable to me.

My thanks also go to all of my colleagues at MARCS especially, Tim Paris, Michael Fitzpatrick, Sonya Prasad, Saya Kawase, April Ching, Julie Beadle and Mandy Visser for creating such a stimulating research environment. Further, my thanks also go to Benjawan Kasisopa for being a wonderful friend, sister and neighbour. I would also like to especially thank Daniel Hochstrasser and Simone Simonetti for your friendship, the great memories and laughs, and for volunteering yourselves for my pilot disgust experiments.

To my parents, Lee Yuen Lin and Chong Ying Keong, thank you for your patience and understanding throughout this candidature. I am very grateful to you for always being supportive of me and the decisions that I make. To my "golden" brothers, Chee Hoong, Chee Weng and Chee Foong, thank you for standing by me through thick and thin. To Mashi and my little nephew and nieces, Clarissa, Tara, Kalden and Sara, thank you for being my source of positivity and energy.

Finally and most importantly, I would like to thank my wife, Yvonne Leung. Your understanding, care, love and support have given me the mental strength to keep going. None of this would have been possible without you.

Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

[redacted]

……..

Abstract

This thesis aims to investigate how spoken expressions of emotions are influenced by the characteristics of spoken language and the facial emotion expression. The first three chapters examined how production and perception of emotions differed between Cantonese (tone language) and English (non-tone language). The rationale for this contrast was that the acoustic property of Fundamental Frequency (F0) may be used differently in the production and perception of spoken expressions in tone languages as F0 may be preserved as a linguistic resource for the production of lexical tones. To test this idea, I first developed the Cantonese Audio-visual Emotional Speech (CAVES) database, which was then used as stimuli in all the studies presented in this thesis (Chapter 1).

An emotion perception study was then conducted to examine how three groups of participants (Australian English, Malaysian Malay and Hong Kong Cantonese speakers) identified spoken expression of emotions that were produced in either English or Cantonese (Chapter 2). As one of the aims of this study was to disambiguate the effects of language from culture, these participants were selected on the basis that they either shared similarities in language type (non-tone language, Malay and English) or culture (collectivist culture, Cantonese and Malay). The results showed that a greater similarity in emotion perception was observed between those who spoke a similar type of language, as opposed to those who shared a similar culture. This suggests some intergroup differences in emotion perception may be attributable to cross-language differences.

Following up on these findings, an acoustic analysis study (Chapter 3) showed that compared to English spoken expression of emotions, Cantonese expressions had less F0 related cues (median and flatter F0 contour) and also the use of F0 cues was different. Taken together,

these results show that language characteristics (n F0 usage) interact with the production and perception of spoken expression of emotions.

The expression of disgust was used to investigate how facial expressions of emotions affect speech articulation. The rationale for selecting disgust was that the facial expression of disgust involves changes to the mouth region such as closure and retraction of the lips, and these changes are likely to have an impact on speech articulation. To test this idea, an automatic lip segmentation and measurement algorithm was developed to quantify the configuration of the lips from images (Chapter 5). By comparing neutral to disgust expressive speech, the results showed that disgust expressive speech is produced with significantly smaller vertical mouth opening, greater horizontal mouth opening and lower first and second formant frequencies (F1 and F2).

Overall, this thesis provides an insight into how aspects of expressive speech may be shaped by specific (language type) and universal (face emotion expression) factors.

List of Publications

Chong, C. S., Kim, J., & Davis, C. (2018). Disgust expressive speech: the acoustic consequences of the facial expression of emotion. *Speech Communication*, 98, (pp. 68-72).

Davis, C., Chong, C. S., & Kim, J. (2017). The effect of spectral profile on the intelligibility of emotional speech in noise. In the proceedings of *INTERSPEECH* 2017, Stockholm, Sweden (pp. 581-585).

Chong, C. S., Kim, J., & Davis, C. (2016). The sound of disgust: how facial expression may influence speech production. In the proceedings of *INTERSPEECH* 2016, San Francisco, USA (pp. 37-41).

Chong, C., Kim, J., & Davis, C. (2015). Exploring acoustic differences between Cantonese (tonal) and English (non-tonal) spoken expressions of emotions. In the proceedings of *INTERSPEECH 2015*, Dresden, Germany (pp. 1522-1526).

Chong, C. S., Kim, J., & Davis, C. (2015). Visual vs. auditory emotion information: how language and culture affect our bias towards the different modalities. In the proceedings of *AVSP*, Vienna, Austria (pp. 46-51).

Chong, C. S., Kim, J., & Davis, C. (2014). The effect of expression clarity and presentation modality on non-native vocal emotion perception. In the proceedings of the *17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, Phuket, Thailand.

Contents

Introduction

Speech is one of the main modes of human communication. It involves the production and transmission of streams of auditory (speech sounds) and visual (e.g., visible movement of the articulators such as lips and tongue) signals to a listener. The acoustic speech signal can convey a precisely structured message to a listener, while also conveying indexical information about the speaker such as their gender, age (Smith & Patterson, 2005) and information about attitudes and emotion. Of particular interest to this thesis is how emotion information may be communicated via the auditory and visual speech signals. Expressive speech involves the concurrent production of speech and expression of emotions and I am interested in examining how the encoding of emotion information interacts with auditory and visual speech production variables, such as the properties of the spoken language and speech articulation gestures (how we configure our lips and tongue to produce speech sounds).

In the work presented in this thesis, I aimed to demonstrate how language specific characteristics may have an impact on emotion production and perception; and how the concurrent production of facial expression of emotions during expressive speech may impact speech articulation. The first three chapters of this thesis examined how the production and perception of spoken expression of emotions differed between tone and non-tone languages (Cantonese and English). The final two chapters of this examined the idea that in expressive speech, the production of facial expressions may affect how we shape our lips to articulate speech sounds. To this end, I examined how the emblematic facial expression of disgust may affect the configuration of the lips during speech articulation and thereby affect F0 and the first and second speech formant frequencies. The following section provides a background for the thesis in the form of a brief introduction to proposals about the functions of emotion.

What are emotions and how might these influence spoken expression?

The question of what emotions are, is too broad and can get bogged down in semantics; indeed LeDoux (2012) even suggests that scientists should stop using the term emotion. As a solution to this, I adopt the solution that Adolphs and Anderson, (2018, p. 99) recommend, that of providing a functional definition. It has been proposed that emotions serve a number of different functions both within and outside the organisms that experience them (see Keltner & Gross, 1999 for a review). Within an organism, emotions are thought to serve regulatory and adaptive functions and increase chances of survival (see Cannon, 1916; Rozin & Fallon, 1987). External to the organism, emotions are also thought to serve a communicative function (see Hutcherson & Gross, 2011). In what follows I focus on these three basic functions of emotion: regulation, adaptation and communication, and how the first two of these may play a role in shaping how we express ourselves.

Regulation

From a regulatory perspective, it is thought that the physiological states of our body changes in response to the presence of stimuli which may be internal or external to the body, and emotions are the behavioural consequences or complex programs of actions that are triggered to return our physiological condition to its neutral state; to achieve homeostasis (see Damasio, 2019 for an extended discussion of the concept and role of homeostasis). There is, however, not always a one-to-one relationship between our physiological states and our emotions as the same physiological state can give rise to different emotions. It is therefore further proposed that our subjective emotions arise from differences in how we appraise the stimulus and its likely consequences; learning from each appraisal to form the basis of future appraisals (see Scherer, Schorr & Johnstone, 2001). Therefore, although emotions have properties of evolved sets of functions that adapt people to the environment (like reflexes),

emotions are also far more multifaceted and flexible than reflexes, i.e., they are influenced by learning and are maintained overtime.

The idea that emotions can be shaped by learning is an interesting one as it holds some explanatory value as to why there are individual differences in emotion expressions and how culture may influence the way we express ourselves. Each culture provides a set of guidelines, values, expectations or norms that allow fellow members to understand, communicate and predict the behaviours of others. As a member of a society, we acquire societal norms and the so called 'display rules' that guide how to react in different situations, e.g., where, and with whom our feelings may be expressed, and how these feelings we may be displayed. Furthermore, as individuals develop within a culture, they learn about the abstract concepts, values and complex emotions that are not tied to the regulation of basic physiological needs. These values and display rules can shape how we appraise our surrounding (whether an action is culturally appropriate) and how we manage and express our emotions depending on the social situation (Scherer, Schorr & Johnstone, 2001). As an example, the concept of 'malu' in the Malay culture is an abstract emotion that involves concepts such as shame, embarrassment and shyness. Understanding the detailed concept represented by malu is a social good, as it enables an individual to be aware of her/his place within a social order, guiding the individual on how to act in a socially appropriate manner (Collins & Bahar, 2000).

Studies on cross-cultural emotion perception have generally found an in-group advantage effect where expressions of emotions are most accurately recognised when they are produced and perceived by members of the same cultural background (see the meta-review by Elfenbein & Ambady, 2002). As an example, in a study examining how Japanese and American participants recognised facial expression of emotions, it is found that when presented with facial expressions of emotions produced by American or Japanese actors, the

American participants were more accurate at recognising facial expressions produced by American actors and vice versa, the Japanese were more accurate at identifying the expressions of Japanese actors (Matsumoto 1992, 1997).

In interpreting these results, researchers have typically invoked high level explanations such as cultural differences (Matsumoto 1992; 1999; Elfenbein, 2013; Elfenbein & Ambady, 2003). For example, individualistic cultures which is exemplified by many Western European cultures is defined as cultures that foster the development of independent construals of self (Markus & Kitayama, 1991), favour personal goals over in-group goals (Yamaguchi, 1994), encouraged rationality and interpersonal exchange (Kim, Triandis, Kitagitcibasi, Choi, & Yoon, 1994) and place more importance on attitudes as relatively important determinants of behaviour. In contrast, collectivistic cultures which are exemplified by the Japanese as well as in other Asian, African, Latin-American and many southern European cultures places emphasis on interdependent selves and relationships, in-group goals and norms as determinants of behaviour (also see Matsumoto, Yoo, & Fontaine, 2008). Thus, in contrast to individuals from individualistic cultures (e.g., North American), those from collectivistic cultures (e.g., Japanese) may inhibit the expression of negative emotions in an effort to promote social harmony and social connectedness (Matsumoto 1992; 1999; Elfenbein, 2013; Elfenbein & Ambady, 2003). These differences in expressive mannerisms and one's unfamiliarity with them may affect cross-cultural emotion recognition accuracy.

There may be, however, other ways in which the production and perception of emotions differ across groups without the need to invoke high level explanations such as cultural display rules. In Chapters 1 to 3 of this thesis, I examined how our expression and perception of emotions may vary as a function of the characteristics of spoken language. The rationale for this proposal is discussed in a subsection of this introduction (below). The general idea here is that differences in how the acoustic property of Fundamental Frequency (F0) is used

in tone and non-tone languages may lead to differences in how these language speakers produce and perceive verbal expressions of emotions.

Adaptation

A second proposed function for emotion that is relevant to this thesis concerns adaptation; specifically, where certain emotions and facial expressions are claimed to be behavioural dispositions that have survival value (Rozin & Fallon, 1987). As an example, actions associated with the emblematic facial expression of disgust, i.e., the narrowing of the eyes, the wrinkling of the nose, the closure and retraction of the lips, and at times, tongue extrusion are thought to serve a functional purpose to reduce risk of infection from contact with contaminants (Rozin & Fallon, 1987; Curtis, De Barra, & Aunger, 2011). In the event that a potential contaminant makes it into the mouth cavity (e.g., biting into an apple to find half a worm), the contaminant can still be expelled through the gag reflex or tongue extrusion.

While purposeful actions such as nose wrinkling and lip closure mainly hold intrinsic value within the acting organism, these actions may also be used as cues for an observer to infer the internal mental state of the acting organism (Ekman, 1992). Over time, as communication and cooperation became increasingly important for survival, adaptive cues which are associated with adaptive functions may have transformed in both form and function into signals of communication, i.e., facial expression of emotions (Shariff & Tracy, 2011; Chapman, Kim, Susskind, & Anderson, 2009; Eibl-Eibesfeldt, 2017; Ekman, 1992).

The idea that facial expressions of emotions serve adaptive functions has implications for how we produce expressive speech. Emblematic facial expression of emotions in general, and disgust in particular, involve changes to the configuration of the articulatory organs, e.g., closure of the mouth and tongue extrusion. As these configurative changes have survival value, these are likely to be ‘in built’ and reflexive actions which take priority over other actions such as aspects speech articulation (e.g., producing clearly pronounced or enunciated

utterances). Given that changes to the size of the mouth opening aperture and tongue position are likely to affect the acoustic properties of speech such as the formant frequencies (Lindblom & Sundberg, 1971), Chapters 4 and 5 of this thesis examined how the facial expression of disgust affects speech articulation.

The current work

In positioning this thesis within the context of other emotion research, it is worth noting that the literature on expressive speech (both general and cross-cultural research) is relatively sparse when compared to that on facial expressions. The majority of studies in emotion research (including all of the ones reviewed above) have predominantly focussed on facial expression of emotions and used static snapshots of emblematic facial expression of emotions as stimuli. While the use of static stimuli may have sufficed for examinations of facial expressions[1], emotion expressions in general, and spoken expressions in particular, are dynamic processes. Hence by examining emotion expression through the lens of speech communication, a significant feature of this thesis is that it not only addresses a gap in the current literature, but also provides a more ecologically valid investigation of two of the most important elements of human communication – emotion and speech.

Besides the need for more ecological stimuli, our understanding of expressive speech has also faced a challenge due to the large number of possible acoustic properties to be examined. Unlike how facial expressions of emotions can be characterised by a relatively limited combination of the activation of different facial muscles, in spoken expressions there is a much larger array of acoustic properties that can be examined and it is unclear precisely which of these may be correlates of emotion expression. In fact, it has been claimed that research on expressive speech has yet to identify any emotion specific acoustic profiles of

[1] Reliance on static stimuli was also due to the technical limitations in earlier research.

emotions (Scherer, 2013). By examining how facial expressions and language characteristics play a role in shaping how we express and perceive emotions, a second feature of this thesis is that it adds to our understanding of what language specific factors and physiological mechanisms (facial expressions) may shape the acoustic profile of expressive speech. The following section details the rationale and structure of the work presented in this thesis.

Language specific differences

The notion that aspects of a language may have an impact on how we produce and perceive emotions is not a novel one. For example, in a perception study examining the accuracy rate of participants from different countries at identifying verbal expressions of emotions produced by German speakers, it was observed that the accuracy of recognising spoken expressions of emotions produced by German speakers deteriorated with increasing dissimilarity between the participants' native language and German (Scherer, Banse, & Wallbott, 2001). In this study, two spoken sentences (constructed using pseudo-words) produced by two German actors in neutral, anger, joy, sadness, fear and disgust (disgust was dropped from the study due to poor recognition accuracy in a pilot stimuli evaluation check) were presented in an emotion identification experiment to participants from nine countries; Germany, Switzerland, Great Britain, Netherlands, United States, Italy, France, Spain and Indonesia.

The study found that recognition accuracy was the highest among the German participants, followed by the French speaking Swiss participants, Great Britain, the Netherlands, United States, Italy, France, Spain and Indonesian. It was reasoned that, with the exception of the Swiss participants, the rank order of recognition accuracy may be explained by language differences. That is, spoken expressions of emotions produced by German speakers were most accurately recognised by those who spoke languages of a Germanic origin (German, Dutch and English), followed by the Romance languages (Italian, French and Spanish); with

the lowest accuracy attained by Indonesian participants who spoke Malay, a language which belongs to the Austronesian language family (Scherer, Banse, & Wallbott, 2001).

Further evidence that language interacts with emotion prosody comes from other perception studies which found an in-group advantage effect where the participants performed significantly better if the expression was produced in their native language than in a foreign language (see Pell, Monetta, Paulmann, & Kotz, 2009; and Thompson & Balkwill, 2006). It should be noted that although these studies did not replicate the finding that the accuracy of emotion recognition tracks with language similarity, it was suggested that the in-group advantage effect may be driven by the interference of language specific features when listening to a foreign language (Pell et al., 2009). That is, differences between languages in properties such as segmental inventory, intonation or rhythmic structure could make the basic task of auditory speech processing and/or the processes of extracting salient emotional features from expressions more difficult.

While the idea that the linguistic properties of a language play a role in shaping spoken expressions is an attractive one, this descriptive explanation lacks a theoretical or well-defined mechanistic models that describe how the encoding and decoding of spoken expressions of emotions may be affected by the linguistic characteristics of a language; and why this may have a detrimental effect for cross-language emotion recognition accuracy. Using Scherer and colleagues study (2001) as an example, it is unclear precisely what language specific differences (e.g., segmental inventory? Intonation?) exist between languages of Germanic origin and those of Indo-European origins; how these differences may affect the encoding of emotions in Germanic languages influence and how these differences underpin the detrimental effect it may have on the Indonesian participants' recognition accuracy.

However, if we were to broaden the definition of language similarity beyond considerations of ancestral language families, there is some evidence that there are certain language specific characteristics that can affect how people produce spoken expression of emotions. For instance, in an acoustic analysis study, it was shown that Mandarin speakers used less Fundamental Frequency (F0) than Italian in the production of spoken expressions (Anolli, Wang, Mantovani & De Toni, 2008). The comparison between Mandarin and Italian is interesting as it provides insights into how language specific differences in Fundamental Frequency (F0) use between tone and non-tone languages may interact with emotional prosody.

Tone and non-tone languages are primarily distinguished by how pitch, the perceptual correlate of F0, is used to achieve lexical distinctiveness (i.e., how words are distinguished). In tone languages, words are distinguished through lexical tones which are produced by variations of F0 that are realised on a segment or syllable. A key feature of tone languages is that several tonal contrasts may be produced on a syllable. As an example, a consonant vowel sequence such as [ma] can mean up to four different meanings depending on its F0 contour; it means "mother" when produced in a high level tone (tone 1) and "scold" in a high falling tone (tone 4). In contrast, lexical distinctiveness in non-tone languages such as Italian and English is predominantly achieved through segmental differences; the different permutations of consonant and vowel combinations.

More importantly, the difference in how F0 is used linguistically in tone and non-tone language may have an impact on how F0 is used in the expression of emotions. That is, while change in F0 is identified as one of the main carriers of emotion information in non-tone languages like English (Scherer & Oshinsky, 1977; Juslin & Laukka, 2003), it may be a less useful cue of emotions in tone language as tone language users may restrict the use of F0 cues in the production of spoken expressions of emotions to preserve pitch as a linguistic

resource that conveys semantic meaning (Ross, Edmondson & Seibert, 1986; Anawin, 1998; Wang & Lee, 2015).

Although the comparison of tone and non-tone languages provides some evidence that emotion production and perception may be driven by language specific differences, there are a number of limitations that needs to be addressed. For one, the evidence (as far as I am aware) is limited to only a single tone (Mandarin) and non-tone language (Italian), and only to the acoustic analyses of spoken expressions (Anolli, Wang, Mantovani & De Toni, 2008; see also Wang & Lee, 2015 which examined only Mandarin expressions). That is, it is unclear if the differential use of F0 (restricted in tone languages vs. one of the most important cues in non-tone languages) may lead to systematic differences in how spoken expressions of emotions are perceived which may in turn, lead to a reduction in cross language emotion recognition accuracy.

Furthermore, the intertwined nature of culture and language means that the purported 'language effect' can also be explained by non-linguistic factors like cultural display rules. Language can be considered as a component of culture and the dissimilarity between languages or language families can be conceptualised as a proxy of cultural difference. Using Scherer and colleagues' study (2001) again as example, those who spoke languages from a common language family of origin (e.g., Germanic) are also more likely to live in closer geographical proximity (and hence likely to have greater exposure to the language in practical exchange like trade, etc) and share greater cultural similarities. The lower recognition accuracy of the Indonesian participants may therefore reflect the geographical and cultural distance of these participants from the Germans.

Given the limitations above, Chapters 1 to 3 of this thesis present a series of production, perception and acoustic analysis studies that have the overarching aim of examining how the

use of F0 in emotion production differs between tone (Cantonese) and non-tone language (English) speakers and how these differences may lead to difference in the perception of emotions. Cantonese is a spoken dialect of Chinese that is widely spoken in Guangdong which is a province in the southern region of China. It is the dominant and official language of Hong Kong and Macau and is widely spoken by Chinese communities in South East Asia such as in Vietnam and Malaysia.

I chose to examine Cantonese (by Hong Kong speakers) in this thesis because it is a tone language that has more lexical tones than Mandarin[2]. The larger number of tones suggests the production of lexical tones in Cantonese may require a more nuanced control of F0, and so Cantonese speakers may therefore place a greater restriction on F0 use in the production of spoken expressions of emotions. This may amplify the potential differences in emotion production and perception between tone and non-tone language speakers. English was my choice for the non-tone language as it is a well-studied language both in linguistic and emotion research and hence serves as a good reference for comparison with Cantonese.

It was in regard to my choice of examining Cantonese that I saw the need to develop an extensive Cantonese expressive speech database, one that consists of a sizeable corpus of spoken sentences produced by a representative sample of both male and female speakers in auditory and visual format. To my knowledge such a database is not available; hence the first goal of my thesis was the construction of the Cantonese Audio-visual Emotional Speech (CAVES) database, which is a relatively comprehensive database of multimodal emotion expressions in Cantonese. While many databases of emotion portrayals have typically focussed on a single modality (facial expressions; e.g. the Faces and Radboud databases (Ebner, Riediger & Lindenberger, 2010; Langner et al., 2010), the creation of a multimodal

[2] six lexical tones in Cantonese and four in Mandarin

corpus allows for the examination of the relative importance of visual and auditory cues, and how the interaction between the two affects emotion perception.

The CAVES database consists of portrayals of the six basic emotions; anger, disgust, fear, happy, sad and surprise; and neutral. These emotions were selected as they are widely considered to be universally expressed and recognised across cultures, and are well studied and represented within the literature on cross-cultural research (Ekman, 1982; Izard, 1977, Scherer, Banse, & Wallbott, 2001; Elfenbein & Ambady, 2002; Juslin, & Laukka, 2003; Bänziger, Mortillaro, & Scherer, 2012). Research generated using the CAVES database is therefore well positioned within the current literature of cross-cultural emotion expression and perception and allows for comparison with other similar studies.

In eliciting the emotion portrayals, 10 native speakers of Cantonese, were required to produce the different emotion expressions with the intent of communicating their affective states to an external observer. The CAVES database therefore consists of emotion expressions that are portrayed with a communicative intent, and are not spontaneously elicited, felt, induced or prototypically posed (where the expresser was coached or instructed on how each emotion should be expressed, e.g. Face and Radboud databases; also see Bänziger, Mortillaro, & Scherer (2012) for a review of different emotion elicitation methods). That is, rather than focusing on the subjective experience of the emotion itself or on expressions that percolate from our internal states, I was interested in emotional expressions that are those typically encountered in our daily lives when interacting with different individuals.

The procedure for emotion portrayals in the CAVES database involves the voluntary repetition of the same sentences in different emotion types. The repetition of the same sentences enables the examination of how various acoustic properties of speech vary as a function of emotion type, while controlling for linguistic factors such as co-articulation

(changes in speech articulation due to neighbouring speech sounds) and, particularly for tone languages, tone-sandhi (where a lexical tone may change due to the pronunciation of adjacent words or morphemes). To ensure that the same sentences can be used for each emotion type, semantically neutral sentences were used in this database.

The first section of the Chapter 1 presents in more detail the overall design and development of the database while the second reports the results of a perception study that evaluated the recordings of the CAVES database. While the CAVES database was developed specifically for the purposes of this thesis, I intend to make this database available to the wider research community.

Having developed the CAVES database, I conducted a perception study which aimed to examine how perception of auditory and visual spoken expression of emotions may be influenced by the interaction between the language that the emotion was expressed in (Cantonese and English) and the perceivers' cultural background and native language. This study is presented in Chapter 2. In this study, three groups of participants who spoke different native languages were recruited for this study; Australian English, Hong Kong Cantonese and Malaysian Malay speakers. Given the intertwined nature of culture and language, an innovative aspect of this experiment was the recruitment of the Malaysian participants who represent an interesting contrast group as they share some similarities with each of the other two groups of participants. That is, Malaysians have a similar collectivist culture as the Hong Kong participants while they speak Malay, a non-tone language which in this regard is the same as English. It should be noted that in conducting this study, care was taken to recruit Malay speakers who do not identify themselves to be of Chinese descent and who do not speak nor understand Cantonese.

Chapter 3 presents an acoustic analysis study that is designed to complement the findings of Chapter 2. The aim of this study is to verify the claim that the F0 use is restricted in the production of spoken expressions of emotions in tone languages (Anolli, Wang, Mantovani & De Toni, 2008; Wang & Lee, 2015) and to examine if there are systematic differences in how F0 related cues are used in the production of emotion expressions spoken in Cantonese and English. It should be noted that the studies presented in Chapters 2 and 3 were conducted when the CAVES database was still in the development and evaluation phase. The stimuli that was available for use, especially for the study in Chapter 2. Preliminary analysis of these limited stimuli found that the recognition accuracy for Fear was low and had to be dropped from the experiments.

Facial expressions

While the previous chapters focussed on the auditory domain examining how language specific differences in terms of F0 use affected the production and perception of spoken expressions, Chapter 4 examined the role that facial expressions of emotions play in shaping the acoustics of expressive speech. Here I tested the idea that in expressive speech, the production of facial expressions may interfere with how we shape our lips to articulate speech sounds.

It is well established that changes to the configuration of the lips can alter the resonant properties of the vocal tract; changes to the size of the mouth opening can affect the first formant frequency of speech (F1) (Lindblom & Sundberg, 1971). Considering that there are some prototypical facial expressions such as smiling that chiefly involve the lower half of the face, it is likely that speech articulation and the F1 of speech will be affected by the concurrent production of facial expressions and speech. Indeed, it has been demonstrated that speech produced while smiling has higher formant frequencies compared to neutral speech (Tartter, 1980).

Surprisingly, despite the fact that facial expressions of emotions clearly affect the mouth region, few studies have explored how facial expressions affect lip configuration during speech articulation and its ultimate result on the acoustic profile of expressive speech. I am only aware of two such studies, one which found that disgust expressive speech was produced with significantly more advanced jaw, nose wrinkling, upper and corner lip raising, and lowering of the larynx (Bailly, Bégault, Elisei & Badin, 2008) while the other noted that that vowels are produced with minimal mouth opening, maximal spreading and retraction of the lips, and negative right vertical asymmetry when produced in disgust (Caldognetto, Cosi, Drioli, Tisato & Cavicchio, 2004). These studies, however, have not examined how these changes in articulation may be associated with measurable and predictable changes in the formant frequencies of speech. Therefore, in Chapter 4, I present a study examining how the facial expression of disgust affects the configuration of the lips during speech articulation and how changes in the lip configuration may affect the fundamental frequency, and the first and second formant frequencies (F1 and F2) of speech.

I chose to examine disgust because it is argued to have a clear evolutionary underpinning. It is widely held that disgust evolved as a pathogen avoidance mechanism and this has led to claims that the emblematic facial expression of disgust serves the goal of reducing the probability that a contaminant or pathogen may enter our bodies through our oral and nasal cavities (Tybur, Liebermann & Griskevicius, 2009; Rozin, Haidt & McCauley, 2008; Curtis, Aunger & Rabie, 2004; Susskind et al., 2008). Given that the facial expression of disgust serves a functional purpose with respect to the preservation of our wellbeing, the selection pressure for such gestures may take precedence over aspects of speech articulation, hence, the emotion type where facial expression may have the clearest and largest effect on speech articulation.

I used the CAVES database for this study as it was the largest AV database of spoken expressions that was available to me. In addition, the design of the CAVES database was best suited for the purposes of this study as the CAVES consisted of a set of 50 stimuli sentences that was produced in the different emotion types and in neutral. By comparing disgust to neutral expressions produced on a consistent set of sentence stimuli, we were able control for linguistic differences at the sentential level such as coarticulation and tone sandhi.

An algorithm was developed as part of this thesis to perform the visual measure of lip configuration. The development of this algorithm is presented in Chapter 5 which is structured as a methods chapter. Each chapter of this thesis with the exception of Chapter 5 consists of published papers and the manuscripts of these papers were fully reproduced. Each chapter therefore includes its own figure numbering system, references list and reference format that was prescribed by the publishers. While I aim to remain faithful to the published work, minor edits were made in some manuscripts to include additional findings and references that I was unaware of at the time of writing.

Chapter 1. The Cantonese Audio-visual Expressive Speech (CAVES) database

This chapter is about the CAVES database that was used in the studies presented in the subsequent chapters of this thesis. The chapter consists of two sections: The first section presents the overall design and development of the database which was written up as a peer-reviewed proceeding that was presented at the 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA) in 2014. This proceeding is presented in full. The second section presents the unpublished results of a perception study which evaluated the recordings of the CAVES database.

The Cantonese Hearing in Noise Test (CHINT) list of sentences is used in the creation of this database. The CHINT is used under licence from the University of Hong Kong; and House, Ear Institute, and we do not have the permission to reproduce a printed list of the stimuli sentences used[3].

[3] For those who are interested in using the CAVES database, please contact the author of this thesis at chongcheeseng138@gmail.com.

Development of an Audio-Visual Cantonese Emotional Speech Database

Chee Seng Chong, Jeesun Kim, Chris Davis,
MARCS Institute,
University of Western Sydney[4]

Abstract

As part of our research examining the auditory and visual properties of the verbal expression of emotion in tone and non-tone languages, we have created a Cantonese emotional speech database. Currently this database consists of recordings of 10 native speakers of Cantonese as they expressed 50 sentences in the 6 basic emotions (angry, happy, sad, surprise, fear, and disgust). This paper reports on the factors that we took into account in the development of this database. In our view, the construction of such a database will provide an important source of evidence for determining the extent to which tone language users differ from non-tone language users in how they express emotions.

Introduction

Most research on emotion expression and recognition across different cultures has been conducted using visual stimuli (facial expressions) [1], [2]. However, the study of the vocal expression of emotion also has a long history [3]–[5] and recently there has been an increase in studies examining auditory-visual expressions [6]–[8]. Studies on vocal emotion have found that emotional expression in English (a non-tone language) is typically conveyed through prosodic cues or changes in the acoustic properties such as Fundamental Frequency (F0), intensity and duration. Studies have also indicated that language plays an important role in the expression of vocal emotions since accuracy at identifying emotions is influenced by

[4] The University of Western Sydney was rebranded as Western Sydney University in August 2015 after this work was published.

whether the language is expressed in one's own native language or not [9], [10], with the size of this effect depending on the similarity of the test language to their native one [11]. That this effect is stronger for auditory compared to visual expressions, suggests that it is due to spoken language and not culture [12].

In order to further examine the relationship between language and emotion expressions we recently conducted a study that tested how native and non-native speakers of English performed in recognizing 5 of the basic emotions expressed with English vocal expressions [12]. For the non-native speaker group we enlisted a group of native Cantonese speakers. Our rationale for selecting Cantonese was because it is a tone language (having 6 phonemic tones, plus 3 "checked" ones) in which each tone is formed through a variation in F0.

It has been suggested that since F0 variation is used in tone languages to convey semantic information, vocal emotion expressions in these languages (compared to non-tone ones) may utilize a different strategy, one that minimizes the usage of F0 for conveying emotional information [13]. In [13] it was reported that speakers of Mandarin (a tone language with only 4 tones) used less F0 related cues than Italian speakers but more intensity and speech rate cues in expressing negative emotions.

Our results supported those of [9] as we found that the non-native listeners (the Cantonese speakers) had a lower accuracy at correctly identifying the expressed emotions when compared to the native listeners (the English speakers) showing an 'in-group' advantage effect. This native/non-native difference only occurred when the stimuli were presented only in the auditory modality, i.e., there was no difference when visual expression information was provided [12].

This effect linking spoken language and emotion is important for the long-standing debate concerning the universality or cultural specificity of emotion expressions. However, before

conclusions can be drawn, it is essential that the basis of the effect and its reliability be determined. Moreover, to properly complete this study, what is needed is to run an identical experiment but this time with Cantonese expressions as the stimuli and English speakers as the non-native listeners group.

It was in regard to conducting such a study that we saw the need to develop an extensive Cantonese emotion database; one that was based on the collection of a sizeable corpus of spoken sentences and comprised both male and female speakers in auditory and visual (AV) format. This was because, to our knowledge, such an AV emotion database is not available. The bulk of Chinese emotional speech databases are created using Mandarin, and the few that exist in Cantonese do not consist of AV stimuli. Likewise, the majority of these databases have only a small number of speakers (one or two speakers) and/or low numbers of utterances (20 sentences). Furthermore, many of these databases have emotion labels that differ from the 6 'basic' emotions (e.g., emotions such as boredom), thus making comparison with previous studies (including our first experiment) impossible. Some databases also employ professional actors, raising the issue of the naturalness of the speech [14]. Although some databases use semantically biased speech to overcome the issue of acted speech, this method is not useful for our research purposes as these sentences may bias the native listener's responses to the speech and not the emotional content. Moreover, it does not allow us to compare the acoustic properties of different emotion expressions using the same sentences. In order to address the issues above, we decided to create our own database.

Database Design

We aim to create an auditory-visual corpus of Cantonese emotion expressions in uttering sentences that can be used in perception studies and cross-cultural comparisons. Here we targeted the 6 basic emotions plus a neutral expression to serve as a baseline. We chose a set

of semantically neutral sentences so that the 6 basic emotions could be expressed without any semantic conflict. This then would allow for the comparison of different emotions to a baseline state using the same stimulus material.

We also deliberately chose sentences that have a good coverage of the different lexical tones both in the initial and final sentence positions. We plan to examine how the onset and offsets of emotion intonation or prosody changes as a function of the different tones. For example: will Cantonese speakers utilize different acoustic cues when expressing anger (associated with a rising tone) when they have to express sentences that have an onset/offset falling versus rising tones?

Further, given that both auditory and visual information carries emotion information, the creation of an auditory-visual spoken emotion database would allow for the investigation of how face and voice work together in the expression and recognition of emotions [8].

Methods

Participants

Ten native speakers of Cantonese (5 females) who were born and raised in Hong Kong were invited to participate for monetary reimbursement. The average age of the participants was 29.1 years (SD = 4.9).

Materials

All sentences stimuli material were obtained from the Cantonese Hearing In Noise (CHINT) sentences list [15]. Of the 240 sentences available, 50 were chosen to be used in this emotion expression production study on the basis that they had a good spread for the different tones at the initial and final position in the sentences. Here we used a six-tone system to classify our stimuli (see table 1). Table 2 shows the number of sentences with each tone as the initial and final position in all of the sentences in the CHINT list. Given that there are many sentences that start with tone 5 but few that end with it, we selected 50 sentences from the list that have

a good balance of tones at the initial and final positions.

Table 1
Example of tone descriptions using the 6 tone classification system[5]

Tone Number	Tone	Description	Example
1	˥	high level	分*
2	˧˥	mid rising	粉
3	˧	mid level	訓
4	˩˧	low falling	墳
5	˨˩˦	low rising	忿
6	˨˩	low level	份

*Here 'fan' is expressed in the 6 different tones.

Table 2
Properties of the sentences selected from CHINT

	CHINT		50 selected CHINT	
Tones	Initial	Final	Initial	Final
1	36*	61	8	12
2	47	45	8	9
3	38	36	8	7
4	19	46	8	8
5	86	7	14	5
6	14	45	4	9

*The numbers in each cell represent number of sentences with the indicated tone at the initial and final position in each sentence.

As the original CHINT sentences were developed to be used for hearing in noise tests, they included a list of parenthesized words that could be substituted and considered as a correct answer in speech identification in noise paradigm. For example, in this sentence:

教授(就快/就嚟)去美國做研究[6],

[5] The translation of the Chinese words in this table from tone 1 to 6 are: point, noodles, discipline, grave, angry, and portion

[6] In English, this sentence reads "The professor will soon be travelling to America to conduct research".

both (就快/就嚟) have the same meaning, “soon”, so for the purposes of this database, we decided to use the second pair of characters 就嚟 to maximise the number of different tones within that sentence. To illustrate this, the original sentence represented as tones would be:

3,6,(6,3/6,4),3,5,3,6,4,3

So, to balance out the ratio of tone 3 to tone 4 characters we picked the second pair of characters resulting in a 10 character sentence with these tones -

3,6,6,4,3,5,3,6,4,3

The same strategy was used for all the selected sentences resulting in each sentence having only ten characters. The motivation behind this selection method is to obtain emotional utterances that consist of a good balance of all 6 tones. As was mentioned, vocal expressions for some emotions in English is carried by pitch properties, as such it is interesting to look at how pitch changes in Cantonese especially on the different tones such as a high level pitch contour in tone 2 or low/falling in tone 4.

Recording Setup

Participants were seated in front of a 20.1” LCD video monitor (Diamond Digital DV201B) that is used to present the stimulus sentences to the participant. Directly above the monitor is a video camera (Sony NXCAM HXR-NX30p) where participants are requested to fixate at prior to expressing the sentences. The videos were recorded at 1920 x 1080 full HD resolution at 50 fps. To capture participants’ utterances a microphone (AT 4033a Transformerless Capacitor Studio Microphone) was placed about 20 cm away from the participants’ lips and out of the field of view of the camera (see Fig. 1 for a picture of the setup). Audio captured using the microphone was fed into the Motu Ultralite mk3 audio interface with FireWire connection to a PC running CueMix FX digital mixer and then to

Audacity which captured the sound at a sampling rate of 48000kHZ. This audio feed as well as video feed from the video camera was monitored by the experimenter outside of the booth who provided the participants with feedback as well as displaying the next sentence on the monitor in front of the participants.

The camera, screen and microphone heights were adjusted to suit each participant so that only the head from the top part of the shoulder was recorded. All participants did a short trial session of three utterances to determine the best gain values for the microphones. Once an acceptable level had been achieved, the recording began and each recording session was blocked by emotion type with the order of the production of type randomized across participants. Participants were given a break after the successful production of every 25 sentences. Through the course of the recording, the experimenter did not interfere, guide or give examples as to how each emotion type should be expressed to avoid demand bias.

Production of Emotion

Participants were invited to take part in the study and were briefed regarding the recording before signing a consent form indicating that recordings could be made public. All participants were instructed to be as natural as possible in how they expressed themselves and were asked to produce the emotions with the intent of communicating their emotional feelings to an observer. Therefore, although the recordings were not “spontaneous” emotions, the participants did strive to express each emotion as if she/he was conveying emotional information to another person. That is, rather than focusing on the experience of the emotion itself, we were interested in emotional expression; the signals that people give to others to express emotion. Such expressions are those typically encountered in our daily lives when interacting with different individuals. In other words, although these portrayed emotion signals may be variable and may differ from emblematic fully fledged expressions that innervate all the facial action units ascribed to a particular expression [16], they were

nevertheless produced to convey emotion. This production method persevered one of the aims of the research project which was to examine how variation in the expression of emotion is tolerated by perceivers.

During the recording sessions, the stimuli sentences were displayed one at a time in a random order on the computer monitor and the participants then produced the utterances when ready. Participants were given feedback via the screen if they had to repeat the sentence (e.g., they misread the sentence or did not fixate on the camera while producing the expressions). Participants were also given three practice trials prior to the start of each emotion block and asked to put themselves in the mode of expressing the emotion. As was mentioned, the emotions to be expressed were blocked, so participants would produce all 50 sentences expressing the same emotion giving a total of 350 sentences per speaker (50 sentences x 7 (six emotions plus neutral).

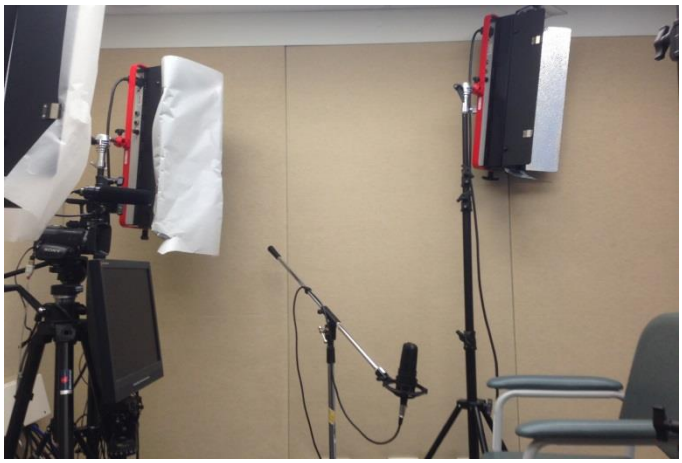

Fig. 1. The setup in the recording booth showing the camera, screen microphone, lighting and participants' seat.

Data Segmentation

Unless otherwise stated, all of the steps described below were done using scripts written in Matlab [17]. The video files were cropped to a more manageable size of 1000 x 1000. In cropping there were two methods available to us: First is dynamic cropping where we have an algorithm that tracks lip movements and crops a set area surrounding the lips, resulting in video clips that are always centered on the participants' lips regardless of head movements. The second method, a static cropping method, is the standard method of cropping a predefined area of the clips. Here we chose to use the standard cropping method to capture all of the head movements made by the participants (see Fig.2 for an example of a cropped video). This is because rigid head motion carries emotion information, which is a variable of interest to us [15].

Once cropped, the audio tracks of the video recordings were replaced with the higher quality audio recordings that were captured separately from the video. To reduce processing times, a copy of the high quality audio recordings was down sampled to 16 kHz and filtered using a high-pass filter of 100 Hz. Noise shaped dither was also applied. The speech and silent segments of the down sampled recordings were automatically annotated using the Audio Segmentation toolkit [18]. The down sampled audio recordings were then discarded and the annotations were manually checked and exported as PRAAT textgrids using the MTRANS program implemented in Matlab [17], [19], [20]. The audio track of the video recordings was replaced with the high quality audio recordings before they were segmented into individual sentence stimuli using the PRAAT textgrid timestamps with an additional buffer of 500 ms before and after the utterance.

Fig.2. A single frame extracted from video clip (high-pass filtered) to show the extent to which the video was cropped

Data Organization

All segmented video clips were then kept in their original format (.MTS) without compression to keep them at the highest quality possible with the unprocessed video and audio recordings preserved. These videos can then be compressed and converted to any lossless format to suit the purposes of other experiments. All segmented clips were labeled by speaker ID, emotion type and then by sentence ID.

Current Progress and Verification

The recordings of 10 speakers have been completed. The recordings are currently being segmented and will be validated in a perception experiment[7].

Conclusion

In summary, as part of our research project, we require the use of an audiovisual Cantonese expressive speech database but due to the dearth of such databases, we have created a Cantonese emotion expression speech database. This paper documents the steps involved and the considerations made in the creation of our database. While there is still much to do, the

[7] See next section for validation study.

recording, segmentation and cataloguing stages are complete and we have begun the verification stage. We plan to conduct acoustic and visual analyses of the stimuli of each speaker in order to add a profile of these characteristics to the database.

Acknowledgment

The creation of this database is supported by the MARCS Institute scholarship. The authors would like to thank the House Ear Institute for the use of the Cantonese Hearing in Noise Test sentences list and Dr. Vincent Aubanel for the Matlab scripts used for the segmentation of the recordings.

References

[1] P. Ekman, W. Friesen, M. O' Sullivan, I. Diacoyanni-Tarlatzis, R. Krause, T. Pitcairn, K. Scherer, A. Chan, K. Heider, W. LeCompte, P. Ricci-Bitti, and M. Tomita, "Universals And Cultural Differences In The Judgment Of Facial Expressions of Emotion," *J. Pers. Soc. Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.

[2] D. Matsumoto and P. Ekman, "American-Japanese cultural differences in intensity ratings of facial expressions of emotion," *Motiv. Emot.*, vol. 13, no. 2, pp. 143–157, 1989.

[3] P. N. Juslin and P. Laukka, "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," *Emotion*, vol. 1, no. 4, pp. 381–412, 2001.

[4] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: different channels, same code?," *Psychol. Bull.*, vol. 129, no. 5, pp. 770–814, 2003.

[5] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *J. Pers. Soc. Psychol.*, vol. 70, no. 3, pp. 614–36, 1996.

[6] G. Pourtois, B. De Gelder, A. Bol, and M. Crommelinck, "Perception of Facial Expressions and Voices and of Their Combination in The Human Brain," *Cortex*, vol. 41, pp. 49–59, 2005.

[7] O. Collignon, S. Girard, F. Gosselin, S. Roy, D. Saint-Amour, M. Lassonde, and F. Lepore, “Audio-visual integration of emotion expression.,” *Brain Res.*, vol. 1242, pp. 126–35, 2008.

[8] J. Kim and C. Davis, “Perceiving emotion from a talker: How face and voice work together,” *Vis. cogn.*, vol. 20, no. 8, pp. 902–921, 2012.

[9] H. A. Elfenbein and N. Ambady, “Is there an in-group advantage in emotion recognition?,” *Psychol. Bull.*, vol. 128, no. 2, pp. 243–249, 2002.

[10] H. A. Elfenbein and N. Ambady, “On the universality and cultural specificity of emotion recognition: A meta-analysis.,” *Psychol. Bull.*, vol. 128, no. 2, pp. 203–235, 2002.

[11] K. Scherer, R. Banse, and H. G. Wallbott, “Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures,” *J. Cross. Cult. Psychol.*, vol. 32, no. 1, pp. 76–92, 2001.

[12] C. S. Chong, J. Kim, and C. Davis, “The effect of expression clarity and presentation modality on non-native vocal emotion perception,” in *The 17th conference of the Oriental chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment /CASLRE (Conference on Asian Spoken Language Research and Evaluation)*. Phuket, Thailand: IEEE, 2014.

[13] E. D. Ross, A. E. Jerold, and G. Seibert, “The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice,” *J. Phon.*, vol. 14, pp. 283–302, 1986.

[14] J. Wilting, E. Krahmer, and M. Swerts, “Real vs acted emotional speech,” in *Interspeech*, 2006, pp. 805–808.

[15] L. L. N. Wong and S. D. Soli, “Development of the Cantonese Hearing In Noise Test (CHINT).,” *Ear Hear.*, vol. 26, no. 3, pp. 276–89, 2005.

[16] R. Reisenzein, S. Bordgen, T. Holtbernd, and D. Matz, “Evidence for strong dissociation between emotion and facial displays: the case of surprise.,” *J. Pers. Soc. Psychol.*, vol. 91, no. 2, pp. 295–315, Aug. 2006.

[17] MATLAB 13.0, The MathWorks, Inc., Natick, Massachusetts, United States.

[18] “Audio Segmentation Toolkit.” [Online]. Available: https://gforge.inria.fr/projects/audioseg. [Accessed: 20-May-2014].

[19] M. Cooke, V. Aubanel, and M. A. Piccolino-Boniforti, “M TRANS: A multi-channel, multi-tier speech annotation tool,” 2011. [Online]. Available: http://www.laslab.org/tools/mtrans/. [Accessed: 17-Jun-2014].

[20] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program],” 2014. [Online]. Available: http://www.praat.org/. [Accessed: 14-May-2014].

Evaluation of the CAVES database

This study examined how well emotions are perceived from the recordings of the CAVES database through an emotion identification experiment. The primary aim of this study was to gauge the validity of the produced spoken emotion in terms of the degree of concordance between the perceived emotion expression and the intended emotion expressed by the speaker. For this, the study examined how accuracy rate and the distribution of response frequencies in recognising emotion from each spoken expression of the database varied as a function of emotion type and presentation modality. The data was also validated in terms of the extent to which the response accuracies and response distributions are similar to the findings of other emotion perception and evaluation studies in the literature.

The secondary objective of this study was to examine speaker and item specific information which can be used as a reference for the selection of stimuli used in subsequent chapters of the thesis. For this, variability in the accuracy rates for at the speaker and sentence stimulus level were examined.

Methods

Stimuli

All expressive speech recordings (50 sentences x 6 basic emotions x 10 speakers = 3,000) were used as stimuli. These recordings were presented in three modality conditions, auditory-visual (AV), visual-only (VO) and auditory-only (AO), resulting in a total of 9,000 stimulus items.

Participants

A total of 41 participants (18 males and 23 females, mean age = 24.7 years, SD = 4.5) took part in this study for a small payment. All participants were native speakers of Cantonese

who were born and raised in Hong Kong. The majority of the participants were recruited through word of mouth and snowball sampling procedures.

Design and Procedure

Due to the large number of stimuli, each participant took part in multiple sessions. Each session was conducted on a separate day and consisted of a total of 900 trials (50 sentences x 3 presentation conditions x 6 emotions) from a random selection of either male or female only speakers. Prior to the start of the experiment, all participants agreed to participate in at least 5 sessions of the study. At the end of the 5th session, participants were given the option to continue participating in the study for an additional session up to a maximum of 10 sessions where all recordings would be rated. The majority (33) of the participants completed more than 5 sessions; 15 participants completed 10 sessions while 2 participants completed less than 5 sessions. Each participant completed an average of 8 sessions.

Given the nature of the experimental design, unequal numbers of data points were generated for each item. Nevertheless, all items were rated by a minimum of 29 participants. The stimuli were presented using DMDX, a free stimuli display program (Forster & Forster, 2003). The program was loaded on a 15.6 inch laptop (Lenovo T520) that is connected to an EDIROL UA-25ex soundcard with Senheisser HD550 headsets. Participants were tested individually in sound-attenuated IAC booths at Western Sydney University.

With no time limit imposed, participants could proceed at a pace that they were comfortable with. Participants were given a 5-minute break every 150 trials and reimbursed for their time at the end of each session.

Analysis

The first set of analyses was conducted on the participants' accuracy data for recognising the different types of emotions across the three presentation modalities. Using the findings of other studies as a benchmark, it was expected that 1) bimodal emotion expressions (AV) be

recognised with higher accuracy than unimodal expressions of VO and AO (see Kim, & Davis, 2012); and 2) recognition accuracy would vary as a function of emotion type, i.e., expressions of Happy were expected to be recognised with the highest accuracy while expressions of Fear at the lowest accuracy (Ebner, Riediger & Lindenberger, 2010; Langner et al., 2010; Tanaka et al., 2015; also see Scherer, Banse & Wallbott, 2001 for auditory only expressions). The second and third analyses were mainly descriptive in nature, exploring speaker and item level differences in accuracy scores.

Results and Discussion

Accuracy

Figure D1 shows overall recognition accuracy as a function of presentation modality and Figure D2 overall recognition accuracy as a function of emotion type. Figure D3 show participants' recognition accuracy across the six emotion types for each presentation condition. One-sample t-tests with Bonferroni correction indicate that all emotion types were recognised at above chance accuracies (6 possible response options = chance accuracy of 16.7%) in all presentation conditions.

Linear mixed models were fitted to the data (using the lme4 package in R, Bates, Maechler, Bolker & Walker, 2015) to examine if recognition accuracy varied as a function of presentation condition, emotion type and the interaction between the two. In these models, speaker and sentence were entered as random factors, emotion type and presentation modality as fixed factors, and recognition accuracy as the dependent variable. The null model included only the random factors and the fixed effects were introduced into the model one at a time in a stepwise manner.

When compared to the null model, the second model which included the random factors and a single fixed effect of presentation modality fit the data significantly better than the null

model, $\chi 2$ (2) = 510.65, $p < 0.001$. This suggests that there was a significant main effect of presentation modality on accuracy.

A third model was then fitted to the data. The third model was essentially the second model with the addition of the second fixed effect of emotion type. The third model was a significantly better fit than the second model, which suggests that there was also a main effect of emotion type, $\chi 2$ (5) = 10111, $p < 0.001$.

A fourth model which included the interaction between presentation modality and emotion type was fitted to the model. When compared to the third model, the fourth model was a significantly better fit, which suggests that in addition to the two main effects, there was also a significant interaction effect, $\chi 2$ (10) = 1037.4, $p < 0.001$.

The significant interaction was examined by conducting separate Kruskal-Wallis one-way ANOVAs of presentation modality for each emotion type using a Bonferroni adjusted alpha. Recognition accuracy varied significantly as a function of presentation modality for each emotion type: anger, $\chi^2(2) = 1112.4$, $p < .001$; disgust, $\chi^2(2) = 2107.2$, $p < .001$; fear, $\chi^2(2) = 277.5$, $p < .001$; happy, $\chi^2(2) = 689.78$, $p < .001$; sad, $\chi^2(2) = 169.64$, $p < .001$; and surprise, $\chi^2(2) = 1393.3$, $p < .001$.

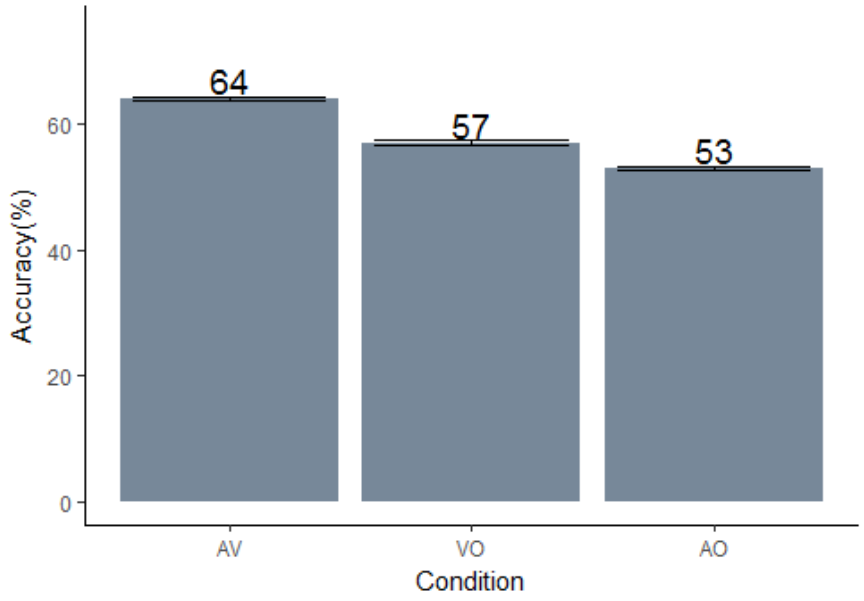

Figure D1. Mean percent accuracy scores across the three presentation conditions. Error bars show standard error.

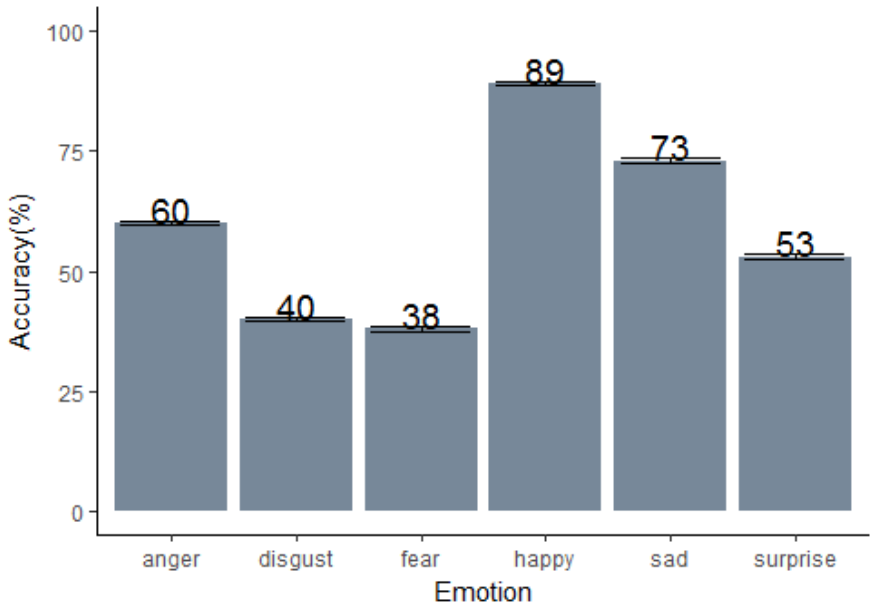

Figure D2. Mean percent accuracy scores across the six emotion types. Error bars show standard error.

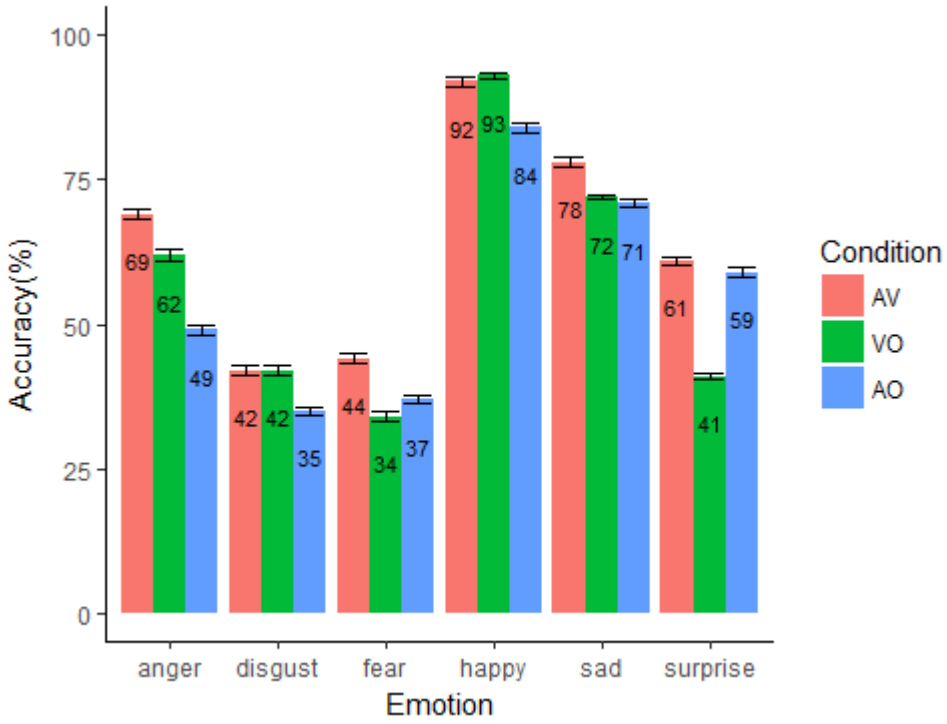

Figure D3. Mean percent accuracy scores for all emotion types in the three different presentation conditions. Error bars show standard error.

Tukey's HSD tests were applied to follow up each significant one-way ANOVA. In general, the patterns in accuracy rates observed in this study were similar to those of the Kim and Davis (2012) study which examined spoken expressions of English that were presented in the three different presentation conditions. Accuracy in the AV condition was significantly higher than both VO and AO conditions for all emotion types except for Disgust and Happy; VO was as accurate as AV for these two emotion types. Comparing the VO to AO condition, accuracy was significantly higher in the VO condition for Anger, Disgust and Happy. This result also closely aligned with the findings of the Kim and Davis (2012) study which found that Anger, Disgust and Surprise (instead of Happy), were recognised at significantly higher accuracy rates in the VO compared to AO condition. Surprise was the only emotion type where accuracy in the AO condition was significantly higher than the VO condition ($p < .001$).

Collapsing across presentation modalities, the accuracy rates observed in this study were similar to those observed in other evaluation studies that have used static images of facial expressions of emotions. For example, Tukey HSD test showed that expressions of Happy were recognised at significantly higher accuracy rates than other emotion expression ($p < .001$); and Disgust and Fear expressions recognised with significantly lower accuracy than all other expressions, $p < .001$. These results were similar to the findings of the evaluation of the Faces, Radboud and Karolinska Directed Emotional Faces database (Ebner, Riediger & Lindenberger, 2010; Langner et al., 2010; Goeleven, De Raedt, Leyman & Verschuere, 2008). The finding that spoken expressions of Fear were recognised with the lowest accuracy was similar to that observed in the Tanaka et al. (2015) study which examined spoken expressions of emotions produced by Japanese and Dutch speakers.

It should be noted that the accuracy rates observed in this evaluation study were lower than those observed for the Faces and Radboud databases. This was most likely due to the following differences in recording procedures. One, the Faces and Radboud databases consisted of static images of facial expressions while the CAVES were dynamic video clips of spoken expressions. Two, expressers in the Faces and Radboud databases were coached to express different facial expressions prior to recording, either by a trained research assistant (Faces) or by two Facial Action Coding System (FACS, Ekman, Friesen, & Hager, 2002) specialists (Radboud), while the speakers in the CAVES database were not given any specific training (apart from three practice trials) and the researcher deliberately kept interference with expresser's performance to a minimum during the recording. These differences meant that the Faces and Radboud databases consisted of clear and unambiguous emblematic facial expressions, while the CAVES consisted to spoken expressions of emotions in face and voice that contained the natural individual differences in speech and expressive mannerisms.

Confusion matrices

Tables 1 to 3 show the confusion matrices for the three presentation conditions. Expressions of Anger were either misidentified as Disgust (AV and AO) or Sad (VO). Expressions of Disgust were either misidentified as Anger (AO) or Sad (AV and VO). Confusion between Anger and Disgust is a common finding observed in evaluations of facial and spoken expressions (Kim and Davis, 2012; Tanaka et al., 2015), It was further observed that negative emotions such as Anger, Disgust and Fear were typically misidentified as Sad; a finding that aligns with previous evaluations of static facial expressions which found Sad to be the most frequently selected response (Goeleven et al., 2008).

Table 1. Confusion matrix for the AV condition.

		Response					
		Anger	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**
Presented Emotion	**Anger**	**69.1**	10.2	3.4	8.5	7.3	1.5
	Disgust	6.4	**41.6**	17.4	5.6	18.4	10.7
	Fear	4.2	11	**43.9**	8	26	7
	Happy	0.4	0.4	0.8	**91.5**	1.7	5.1
	Sad	2.4	6.5	4.8	7.8	**78**	0.4
	Surprise	1.4	1.7	5.7	27.7	2.5	**61**

Table 2. Confusion matrix for the VO condition.

		Response					
		Anger	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**
Presented Emotion	**Anger**	**62.3**	8.2	5.2	7.2	14.9	2.2
	Disgust	7.8	**42**	19.3	5.2	21.5	4.1
	Fear	9.1	13.9	**33.7**	8.1	27.5	7.6
	Happy	0.6	1.1	0.9	**92.8**	3.5	1.1
	Sad	7.6	7.7	6	5.6	**72.1**	0.9
	Surprise	7.5	4.7	5.8	34.6	6.7	**40.6**

Table 3. Confusion matrix for the AO condition.

		Response					
		Anger	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**
Presented Emotion	**Anger**	**48.7**	29.7	2.4	10.6	5.8	2.9
	Disgust	22.6	**35**	10.9	7.1	13.4	11
	Fear	4.9	6.6	**37.4**	14.7	30.3	6.1
	Happy	2	4.8	1.9	**83.6**	6.3	1.4
	Sad	0.9	6.9	7.2	12.9	**71.1**	1.1
	Surprise	6.4	2.8	4.2	23.9	3.3	**59**

Note. For Tables1 – 3, percent correct emotion identification accuracy is underlined and the most prominent response competitor is highlighted.

Fear was generally misidentified as Sad across all presentation modalities. This was similar to the results reported by Tanaka et al. (2015) and by Banse and Scherer (2001). This is however in contrast to some studies that reported that Surprise is the most likely alternative response (see Goeleven et al., 2008 and Biehl et al., 1997). Expressions of Happy and Fear were generally rarely confused with other emotion types.

Interestingly, expressions of Sad in the AO condition were at times misidentified as Happy, further investigation of the data suggests that this was mainly driven by the stimuli produced by one of the male speakers[8]. Across all presentation conditions, expressions of Surprise were most likely to be misidentified as Happy which is also a commonly observed finding (see Kim and Davis, 2012; Tanaka et al., 2015).

Variability between speakers

Figure D4 shows participants' mean percent accuracy score for identifying emotion expressions that were produced by each of the 10 speakers in the CAVES database. In Figure D4, female speakers were given identifiers that started with 'F' with a number from 1 to 5 to denote each individual speaker. Similarly, males were given identifiers that started with 'M'.

[8] Speaker M5 whose expressions were recognised with the lowest accuracy scores (see Figure D4)

A Kruskal-Wallis test indicated that emotion expressions produced by female speakers were generally recognised at a higher accuracy than male speakers $\chi^2(2) = 172.7$, $p < .001$. This is common finding in the literature of emotion perception studies (for example, see Palermo & Coltheart 2004; Wells, Gillespie & Rotshtein, 2016).

Tukey HSD tests indicated that expressions produced by speaker F2 and F1 were recognized at significantly higher accuracy rates than all other participants ($p < .001$). The difference between F1 and F2 was not significant. Accuracy at recognising the expressions produced with speaker M5 was significantly lower than all the other speakers ($p < .011$).

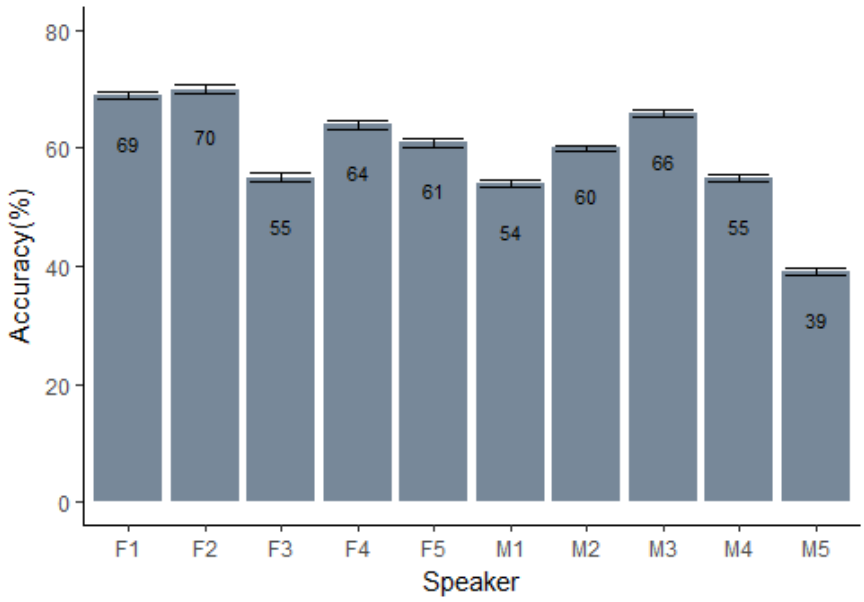

Figure D4. Mean percent accuracy scores for each speaker in the CAVES database with standard error bars.

Item analysis

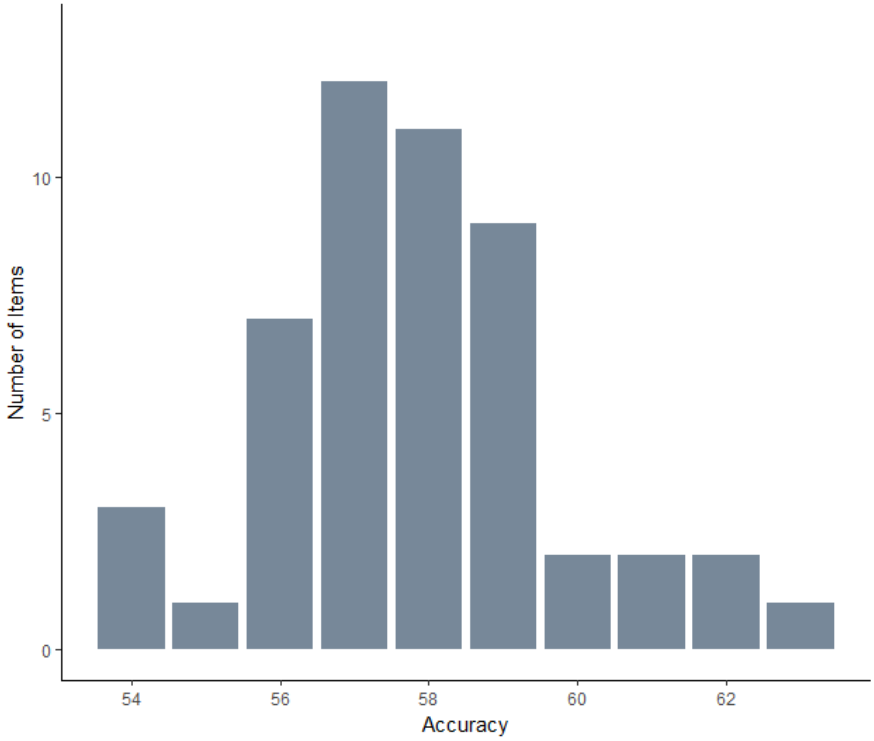

Figure D5. Histogram of item accuracy scores

There were a total of 50 different sentence stimuli that were recorded in the CAVES database. Collapsing across emotion type and speakers, participants recognised all sentence stimuli within the range of 54% and 63%. Figure D5 shows the histogram of accuracy scores. The majority of the items (39 out of 50sentence stimuli) were rated within four percentage points of difference; 39 sentence stimuli were identified within a range of 56% – 59% accuracy rates. The boxplot in Figure D6 shows the outliers and the distribution of scores for all sentence stimuli across each emotion type. The four outliers (sentence stimuli with identifiers of 0302, 0315, 0412, 0510) were not used in subsequent experiments.

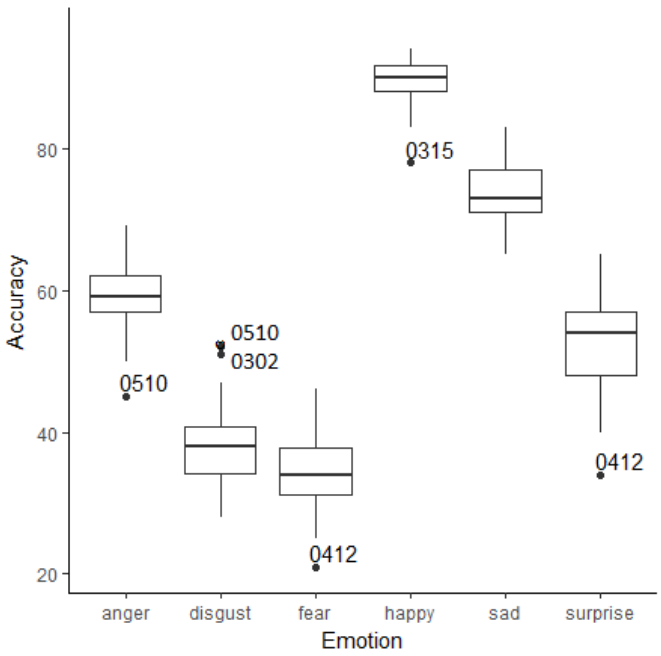

Figure D6. Boxplot showing distribution of accuracy scores for all 50 items across the six emotion types.

In sum, the emotion expressions of the CAVES database can be recognised at above chance accuracy rates; show a significant AV benefit effect, and the patterns of response distribution frequencies are similar to those of other emotion perception and evaluation studies. This confirms the validity of the emotion recordings of our database (Ebner, Riediger & Lindenberger, 2010; Langner et al., 2010; Tanaka et al., 2015; Scherer, Banse & Wallbott, 2001; Goeleven et al., 2008; Biehl et al., 1997; Palermo & Coltheart 2004; Kim & Davis, 2012; and Wells, Gillespie & Rotshtein, 2016). While slight differences with other studies were found in the patterning of the confusion matrices, these differences were likely driven by idiosyncratic individual differences (speaker M5 for example). Cultural or language differences between Cantonese speakers and the expressers examined in the other studies may also have contributed to the discrepancy between results (e.g., English speakers in Kim and Davis 2012; Japanese and Dutch in Tanaka et al. 2015; and German expressers of the Radboud database, Ebner, Riediger & Lindenberger, 2010).

Furthermore, the low variability in accuracy scores across individual items confirmed the reliability of our stimulus selection procedure and suggest that the selected sentences (with the exception of a few outliers) were semantically neutral and did not bias responses to any particular emotion type. In addition, this evaluation study also provided useful information for the selection of stimuli used in the studies presented in the other chapters of this thesis. In summary, we are satisfied that this database was developed as intended and is suitable for the purposes of this thesis. Upon completion of this thesis, we plan to make this stimulus set available to the wider research community.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S., (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal behavior*, *21*(1), 3-21.

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, *42*(1), 351-362.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). Facial Action Coding System: the Manual. Research Nexus, Div. *Network Information Research Corp., Salt Lake City, UT*, *1*, 8.

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska directed emotional faces: a validation study. *Cognition and emotion*, *22*(6), 1094-1118.

Kim, J., & Davis, C. (2012). Perceiving emotion from a talker: How face and voice work together. *Visual Cognition*, *20*(8), 902-921.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. D. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and emotion*, *24*(8), 1377-1388.

Palermo, R., & Coltheart, M. (2004). Photographs of facial expression: Accuracy, response times, and ratings of intensity. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 634-638.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural psychology*, *32*(1), 76-92.

Tanaka, A., Takagi, S., Hiramatsu, S., In't Veld, E. H., & de Gelder, B. (2015). Towards the development of facial and vocal expression database in east Asian and Western cultures. In *AVSP* (pp. 63-66).

Wells, L. J., Gillespie, S. M., & Rotshtein, P. (2016). Identification of emotional facial expressions: effects of expression, intensity, and sex on eye gaze. *PloS one*, *11*(12), e0168307.

Chapter 2. Visual vs. Auditory Emotion Information: How language and culture affect our bias towards the different modalities

The general aim of this chapter was to determine if the perception of the spoken expression of emotions differed between tone and non-tone language speakers. The idea here is that there may be systematic differences in how these speakers perceive emotions due to language specific differences in how F0 is used in the production of expressive speech in tone and non-tone languages.

The first section of this chapter presents the findings of an emotion identification experiment that examined how the perception of spoken expressive speech presented in different modalities (AO, VO and AV) may be influenced by the interaction between the type of language (tone, Cantonese vs. non-tone, English) that the emotion was expressed in and the perceivers' culture and native language. Three groups of participants who spoke different native languages were recruited for this study; Australian English, Hong Kong Cantonese and Malaysian Malay speakers. Given the intertwined nature of language and culture, an innovative aspect of this experiment is the recruitment of the Malaysian Malay participants who represent an interesting contrast group as they share some similarities with each of the other two groups of participants. That is, Malaysians have a similar collectivist culture as the Hong Kong participants while they speak Malay, a non-tone language which in this regard is the same as English.

By examining the similarities and differences in the patterns of responses in how these participants recognise spoken expressions of emotions, I aim to gain an insight into the relative contributions of culture and language (tone vs. non-tone) to the perception of spoken emotions. The general idea is that the impact of language differences on emotion perception will be evidenced by the finding that the Malay participants perceived emotion expressions

more similarly to those who spoke a similar type of language (i.e. non-tone language speakers, English speakers) than those who spoke different types of languages (i.e., Malay and Cantonese). On the other hand, a culture effect will be evidenced by greater similarities in perception between the Malay and Cantonese participants.

In this study, emotional labels were presented in English to all participants. This was done to avoid the introduction of subjective biases and errors associated with the translation of labels from one language to another. As an example, while "anger" is typically translated into Malay as "marah", doing so is inaccurate as "marah" represents a wide range of more complex emotional states such as "offended" and "resentful" (see Goddard 1996 for a review of the difficulties and inaccuracies in translating emotional terms between English and Malay).

The work presented in the first section of this chapter was presented at the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing held in Vienna, Austria in 2015. The proceeding was published as a peer-reviewed proceeding in the ISCA archive[9]. This chapter reproduces the proceeding with some edits (and corrections) made. The second section of this chapter presents the unpublished results of an additional analysis that examined the extent to which the responses of one group of participants matches or agrees with the responses of another group. Here the idea is that if there are systematic differences in how tone and non-tone language speakers perceived emotions, there will be a greater degree of consensus or agreement in what the (non-tone) Malay and English speakers perceive to be the emotion that was expressed in each stimulus item. That is, when responding to each stimulus item in the experiment, the Malay participants may be more likely to choose the same response options as the English speakers than the Cantonese

[9] The proceeding is available online: https://www.isca-speech.org/archive/avsp15/av15_046.html

speakers. In this analysis, the degree of agreement or match in responses between each pairwise group of participants was determined using Fleiss' Kappa.

Visual vs. Auditory Emotion Information: How language and culture affect our bias towards the different modalities

Chee Seng Chong, Jeesun Kim, Chris Davis,
MARCS Institute,
University of Western Sydney

Abstract

This study investigated if familiarity with a language that an emotion is expressed in, affects how information from the different sensory modalities are weighed in auditory-visual (AV) processing. The rationale for this study is that visual information may drive multisensory perception of emotion when a person is unfamiliar with a language, and this visual dominance effect may be reduced when a person is able to understand and extract emotion information from the language. To test this, Cantonese, English and Malay speakers were presented spoken Cantonese and English emotion expressions (angry, happy, sad, disgust and surprise) in AO, VO or AV conditions. Response matrices were examined to see if patterns of responses changed as a function of whether the expressions were produced in their native or non-native language. Our results show that the visual dominance effect for Cantonese and Malay participants changed depending on the language an emotion was expressed in, while the English participants showed a strong visual dominance effect regardless of the language of expression.

Introduction

Compared to the unimodal presentation of emotions, when both facial and auditory expressions are presented together, emotions are typically recognized faster and more

accurately [1, 2, 3, 4]. Although facial expressions typically influence emotion judgments more than auditory expressions (the visual dominance effect) [1, 5, 3, 6], the process of weighing up of emotion information from auditory and visual sources appears to be a flexible one. Typically, when one source of information is degraded (e.g. visual), a greater weight is placed on the complementary source (auditory) [7].

Interestingly, there is evidence to suggest that the weighting of auditory and visual emotion information is not as flexible as has been previously thought. For example, Tanaka and colleagues [8] showed that Japanese perceivers placed a greater importance on auditory emotion cues even when told to ignore the auditory modality. Moreover this tendency to be influenced by vocal cues was greater for Japanese than Dutch participants and occurred regardless of whether the person expressing the emotion was Japanese or Dutch. This is rather surprising as it indicates that the Japanese exhibited a weaker visual dominance effect even when the auditory signal was difficult (i.e., listening to a non-native language can be considered an adverse listening condition).

Tanaka et al. interpreted their findings in terms of Japanese perceivers being more sensitive to context (whereby the face is central to emotion expression and the voice is context). However, a slightly different explanation can be couched in terms of exposure bias based on signal distinctiveness. It is generally acknowledged that there is a Japanese cultural restriction on the overt display of facial emotions [9]. This may result in a bias in which Japanese perceivers would favour the auditory modality over the visual one. Regardless of the exact explanation of the effect, the Tanaka et al. study raised several important considerations. It highlighted cultural differences in emotion perception, and it focused attention on auditory emotion cues. Given the above, the current paper investigated another way that emotion perception may be influenced. Here, we examined how the perception of

auditory-visual emotion expressions may be affected by an interaction between the properties of ones' native language and the language that the emotion is expressed in.

Spoken expressions of emotions carry both linguistic and emotional prosody and there is evidence to suggest that the linguistic properties of speech interact with how emotions are expressed. It is generally found that participants performed significantly better if the expression was produced in their native language than in a foreign language (see [10] and [11]) and that recognition accuracy may deteriorate with increasing dissimilarity between the participant's native language and the language of expression. For example, it was observed that accuracy at recognising spoken expressions of emotions produced by German speakers was the highest among participants who spoke languages of a Germanic origin (Dutch and English), followed by the Romance languages (Italian, French and Spanish); with the lowest accuracy attained by Indonesian participants who spoke Malay, a language which belongs to the Austronesian language family [12].

In interpreting these results, it is claimed that differences between languages in properties such as segmental inventory, intonation or rhythmic structure may interfere with the processes of extracting salient emotional features from the spoken expression. For example, an acoustic analysis study, it was shown that Mandarin speakers used less Fundamental Frequency (F0) than Italian in the production of spoken expressions [13]. The comparison between Mandarin and Italian is interesting as it compares two different types of language, a tone (Mandarin) to a non-tone language (Italian) which are defined by how F0 is used to achieve lexical distinctiveness (i.e., to distinguish words). This comparison also provides insights into how language specific characteristics may interact with emotional prosody.

In tone languages such as Mandarin, variations in Fundamental Frequency (F0) can be used to distinguish the same segments (and used to create lexically distinct items, i.e., words). In

non-tone languages such as Italian, lexical distinctiveness is predominantly achieved through segmental differences; i.e., via permutations of different consonant and vowel combinations. More importantly (for current concerns), change in F0 is one of the main carriers of emotion information in non-tone languages like English (10), and it has been suggested that tone language users may use less variation in F0 in the production of spoken expressions of emotions to preserve the integrity of lexical tones [14, 15, 16].

While the comparison between tone and non-tone languages provides some insight into how language specific characteristics may impact emotion production, to our knowledge, this evidence has been limited to only a single tone (Mandarin) and non-tone language (Italian), and only to the acoustic analyses of spoken expressions [12]. That is, it is unclear if the differential use of F0 (restricted in tone languages vs. one of the most important cue in non-tone languages) may lead to systematic differences in how spoken expressions of emotions are perceived which may in turn, lead to a reduction in cross language emotion recognition accuracy. Furthermore, to-date, we are unaware of any studies that have compared the relative contributions of culture and language to emotion perception.

Given the above, the motivation of the current experiment is straightforward. We propose that with tone and non-tone languages, the weight assigned to the information from auditory and visual emotion signals should vary as a function of the type of language (tone or non-tone) that the target emotion is expressed in. Our hypothesis is that, the size of any visual bias effect (greater weight on visual information) will be greater when the native language of the participant is incongruous with the language that the emotion was expressed in (e.g., non-tone language speakers identifying expressions spoken in a tone language). This is because in such a case, differences in emotion prosody will introduce a degree of uncertainty in the perception of the auditory signal and so make facial information relatively more potent. In contrast, if relative reliance on auditory emotion information is a cultural effect (due to a

sensitivity to auditory context), then similar to Tanaka et al.'s study [8], an auditory bias effect should be observed irrespective of the language the emotion was expressed in so long as the perceiver's culture emphasizes a reliance on context.

In this study, three groups of participants (Australian English, Hong Kong Cantonese and Malaysian Malay speakers) were recruited to take part in an emotion identification experiment. In the experiment they were required to identify the emotion expressed in a spoken sentence that was produced in either Cantonese (tone language) or English (non-tone language). An innovative aspect of this experiment is the recruitment of the Malaysian Malay participants who represent an interesting contrast group as they share some similarities with each of the other two groups of participants. That is, the Malaysians presumable share a similar collectivist culture as the Hong Kong participants, while they speak Malay, a non-tone language which in this regard is the same as English.

Placing the above within the context of our hypotheses, if the cultural bias effect, predisposes a sensitivity to the 'contextual' voice information or face avoidance, the Cantonese and Malay participants would exhibit a greater auditory bias than the Australian participants [8], irrespective of whether the emotion was expressed in English or Cantonese, while the English participants will most likely favour visual cues regardless of the language of expression. Alternatively, if language (tone versus non-tone) differences has a larger effect, then a visual bias might be expected only when the participants were presented emotions that are expressed in the incongruous language. Specifically, the Cantonese participants will show a visual bias effect when identifying English expressions while the English and Malay participants will exhibit the visual bias effect when identifying Cantonese expressions.

Methods

Design

This study used a mixed between-within subjects design where two versions of the experiment were created. One version presented only spoken English expressions as stimuli while the other presented only Cantonese expressions. Participants participated in either the Cantonese or English versions. This was done to prevent participants from developing response strategies based on one language version and using these with the other. In each version, participants were presented trials that were blocked by presentation condition (AO, VO and AV) and the order of presentation was counter-balanced. Within each presentation condition, the trials were blocked by speaker so participants viewed all of the expressions from one speaker before they are presented with expressions from another. The order of speaker presentation was also counter-balanced.

Participants

Cantonese participants: A total of 32 (20 females) native speakers of Cantonese who were born and raised in Hong Kong were invited to participate for monetary reimbursement. 16 (10 females) of them participated in the English version of the experiment. The average age of the participants was 27.5 years.

English participants: A total of 27 (18 females) speakers of Australian English from the University of Western Sydney, Australia participated for course credit. 11 (7 females) of them participated in the English version of the experiment. The average age of the participants was 22.4 years.

Malay participants: A total of 30 (18 females) speakers of Malay from Sunway University, Malaysia participated for monetary reimbursement. 15 (8 females) of them

participated in the English version of the experiment. The average age of the participants was 21.2 years.[10]

The Cantonese and English participants were recruited and tested at Western Sydney University, Australia while the Malay participants were recruited and tested at Sunway University, Malaysia.

Stimuli

All video recordings were edited using FFMPEG [17], VirtualDub [18] and Matlab [19] to create the audio only and visual only stimuli. All video clips were presented at a resolution of 800 x 600 pixels.

English

The stimuli were selected from our laboratory stimuli database. These consisted of five male native speakers Australian English recorded while producing different emotions when uttering sentences selected from the Semantically Unpredictable Sentences [20]. Two speakers and eight sentences were selected and used as stimuli. Two sentence stimuli were selected from a third speaker to be used only in the practice sessions (see [3] for recording procedure).

Cantonese

The stimuli were selected from our laboratory stimuli database[11]. Five male native speakers of Cantonese were recorded while producing different emotions when uttering semantically neutral sentences selected from the Cantonese Hearing In Noise (CHINT) sentences list

[10] It should be noted that in conducting this study, care was taken to recruit Malay speakers who do not identify themselves to be of Chinese descent and do not speak nor understand Cantonese.

[11] CAVES database from Chapter 1.

[21]. Two speakers[12] and eight sentences were selected and used as stimuli. Two sentence stimuli were selected from a third speaker to be used only in the practice sessions (see [22] for recording procedure).

Procedure

Participants were given written instructions and a short practice session prior to the start of the experiment. In the practice session, participants were first presented two video clips (or audio clips depending on the presentation condition) of the speaker uttering a sentence in a neutral expression. These neutral expressions were included to help familiarise the participants with the speaker and acted as a speaker specific baseline against which to judge the emotional expressions. The neutral utterances were followed by 12 practice trials. Each trial consisted of one emotion expression and participants were required to identify the emotion by responding to a five alternative forced choice task using the mouse. The experimental trials were presented once participants completed all 12 practice trials. The researcher remained with the participant during the practice session to ensure that the participants understood the task.

The experimental trials were presented in the same format as the practice trials. The trials were blocked by presentation modality and then by speaker. Participants were always given two sentences in a neutral expression at the beginning of each block. Each block consisted of 40 test trials giving a total of 240 trials (2 talkers x 8 sentences x 5 emotions x 3 presentation conditions) in each version of the experiment. Participants were tested individually in a sound-attenuated lAC booth and stimuli were presented using DMDX [23] on a laptop (Lenovo ThinkPad T520) with the auditory stimuli played through Senheisser HD550 headsets connected to an EDIROL UA-25ex soundcard.

[12] Speakers M2 and M4 of the CAVES database.

Analysis

In quantifying what constitutes a visual or auditory bias effect, studies using the cross-modal bias paradigm like Tanaka and colleagues [8] have looked for a bias to be shown in participant responses to incongruent trials (where the face and voice have been manipulated to express two different emotions). An auditory bias effect is shown if participants judged that the expression was the one expressed in the voice.

In contrast, this study used a more straightforward approach by examining the full response matrices of participants. We did this by subtracting the percentage frequency of participants' responses in the VO from the AV condition, in this way; we can estimate the relative contribution of AO to AV responses. Here we focused on error or perceptual confusion instead of the correct responses because the errors may be less influenced by the AV benefit effect. That is, errors are most likely modality specific, since such errors will not be made if the expressions are recognized as the same emotion in both modalities. Specifically, we looked for increases in auditory-based errors made in the AV condition, because such an error would mean that AO cues were given sufficient weight to influence the outcome of perception in the AV condition. Conversely, if the difference between AV and VO is small, then participants demonstrate a visual bias since auditory information did not contribute to the outcome of AV perception.

Results

English Participants

The confusion matrices were created by subtracting the percentage frequency of responses in the VO from AV condition. A green circle indicates that for the presented emotion, a higher frequency of responses was given in the AV condition, while a clear circle indicates that a higher frequency was observed in the VO condition. To assist interpretation, Figure 1 shows that when presented with Cantonese expressions of disgust, participants in the English

language group were significantly more likely to correctly identify the expression as Disgust (12.1 %), and less likely as Sad (15.4 %) in the AV condition when compared to the VO condition F(1,28) = 15.39, $p < .01$. In other words, the provision of auditory cues in the AV condition has increased the accuracy at identifying Disgust expressions by reducing the confusion with Sad. None of the other changes were significant.

Figure2 shows the English participants responses to English expressions of emotions. None of the differences were significant. The pattern of results in Figures 1 and 2 suggests that the English participants showed a strong visual dominance effect as the provision of auditory cues 1) did not appear to have a significant impact on the pattern of distribution frequencies in the AV condition when compared to VO and 2) did not significantly improve emotion recognition accuracy in the AV condition when compared to VO, Cantonese F(1,19) = 2.25, *ns*, and English F(1,19) = 1.90, *ns (*see Figure 3*)* .

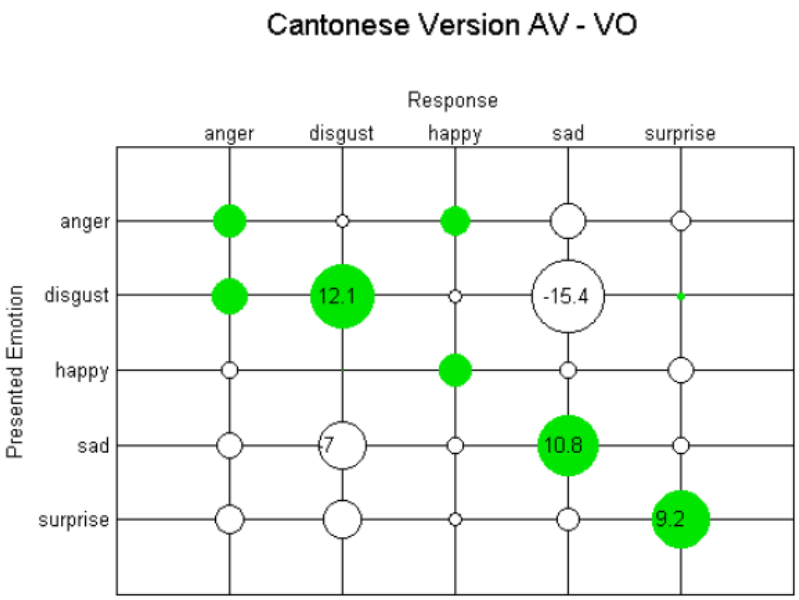

Figure 1. *Confusion matrix showing the differences in the percentage frequency of responses in the AV to VO conditions of English participants who participated in the Cantonese version of the experiment.*

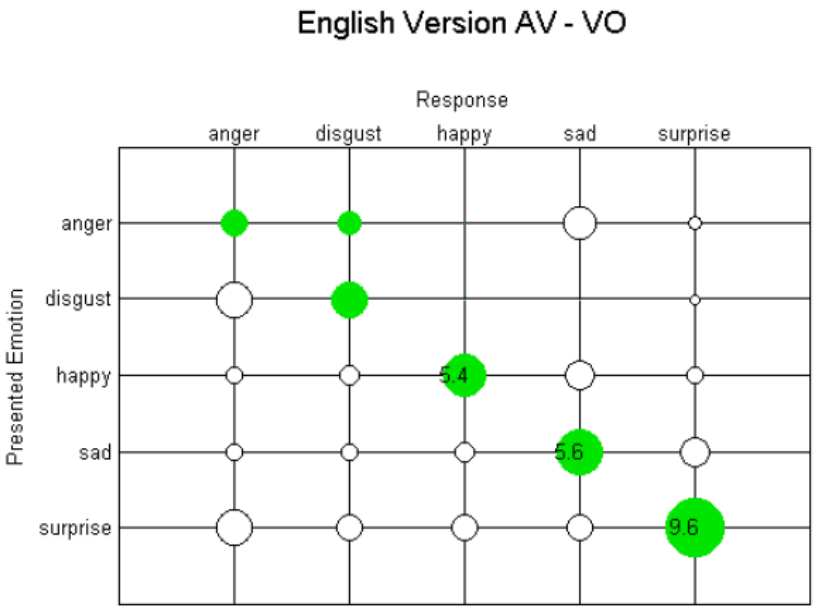

Figure 2. *Confusion matrix showing the differences in the percentage frequency of responses in the AV to VO conditions of English participants who participated in the English version of the experiment.*

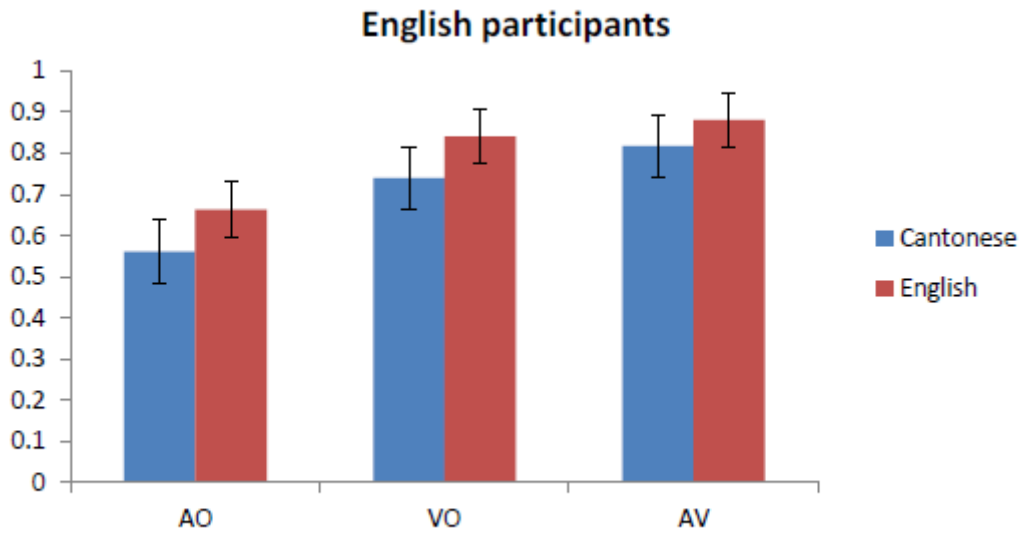

Figure 3. *English participants' accuracy in the English and Cantonese version of the experiment with standard error bars.*

Malay Participants

Figure 4 shows that the provision of auditory information significantly reduced the confusion between, Disgust and Sad $F(1,28) = 16.45$, $p < .001$ and; Surprise and Anger, $F(1,28) = 17.66$, $p < .001$, resulting in a significant increase in accuracy at recognising expressions of Disgust and Surprise, $F(1,28) = 16.98$, $p < .001$ and $F(1,28) = 18.44$, $p < .001$ respectively. Similar to the English participants, expressions of Disgust were frequently misidentified as Sad.

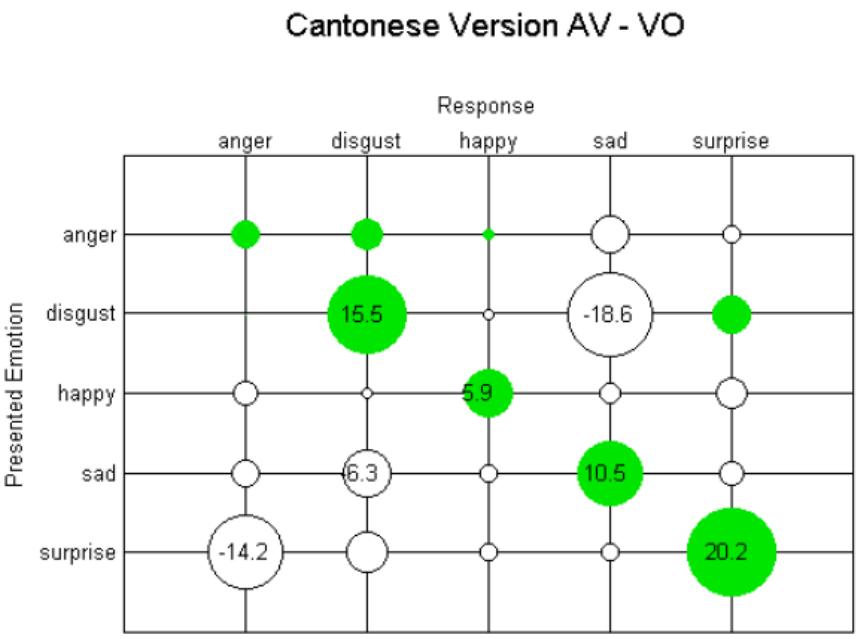

Figure 4. *Confusion matrix showing the differences in the percentage frequency of responses in the AV to VO conditions of Malay participants who participated in the Cantonese version of the experiment.*

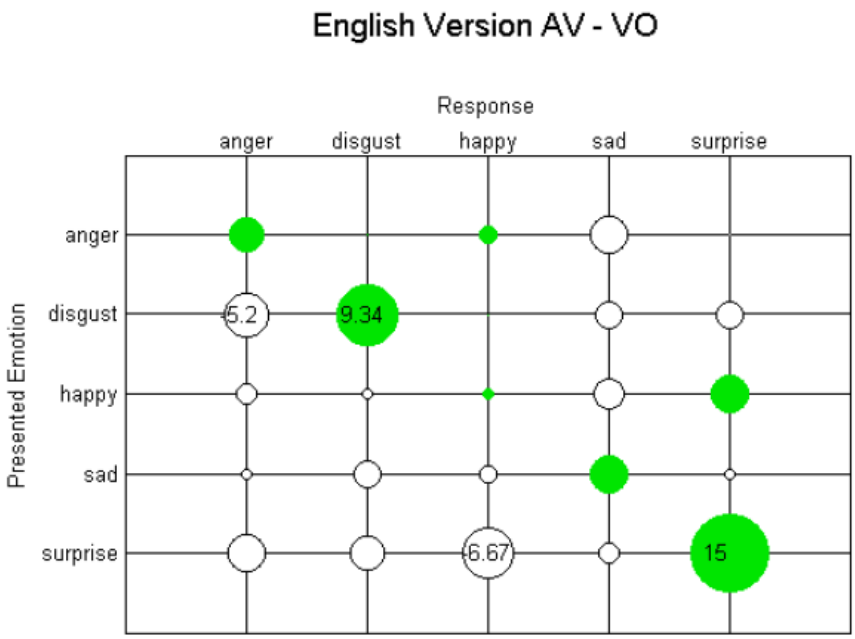

Figure 5. *Confusion matrix showing the differences in the percentage frequency of responses in the AV to VO conditions of Malay participants who participated in the English version of the experiment.*

Figure 5 shows the Malay participants' responses in the English version of the experiment. Although the accuracy at recognising expressions of Surprise was significantly higher in the AV condition, $F(1,28) = 15.65$, $p < .001$, none of the reductions in confusion were significant. Figure 6 shows that there was a significant difference between AV and VO only in the Cantonese version of the experiment $F(1,19) = 4.55$, $p < .05$.

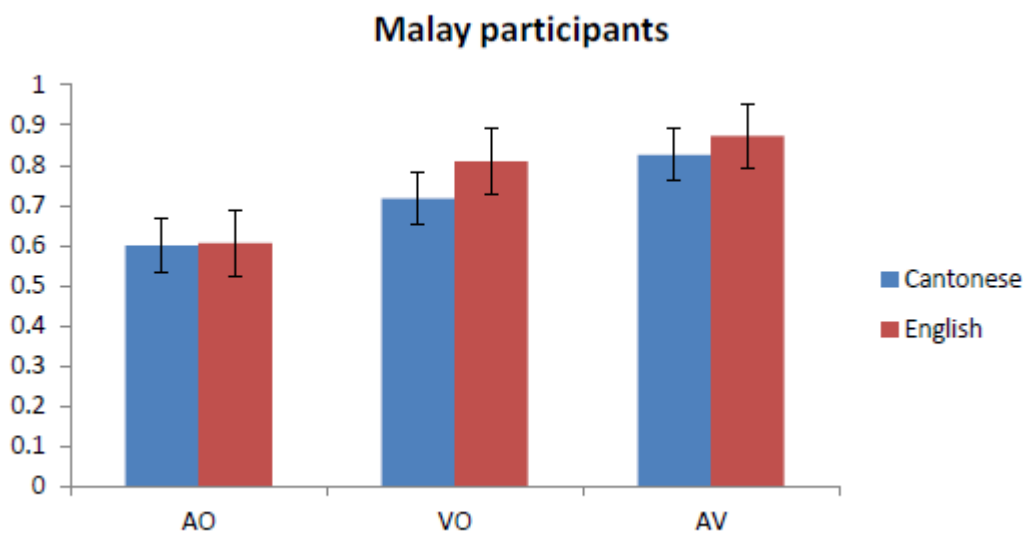

Figure 6. *Malay participants' accuracy in the English and Cantonese version of the experiment with standard error bars.*

Cantonese Participants

Figure 7 shows that the provision of auditory information significantly increased participants accuracy at recognising expressions of Surprise $F(1,30) = 8.03$, $p < .05$ which was mainly driven by a reduction in confusion between Surprise and Anger $F(1,30) = 8.13$, $p < .05$.The confusion between Disgust and Sad was reduced significantly $F(1,30) = 8.54$, $p < .05$. Unexpectedly, the provision of auditory information appeared to have introduced a greater degree of confusion between Disgust and Surprise $F(1,30) = 7.34$, $p < .05$.

While Figure 8 shows that there was an increase in accuracy at recognising Surprise, $F(1,30) = 13.63$, $p < .01$, none of the reductions in confusion were significant. Figure 9 shows the

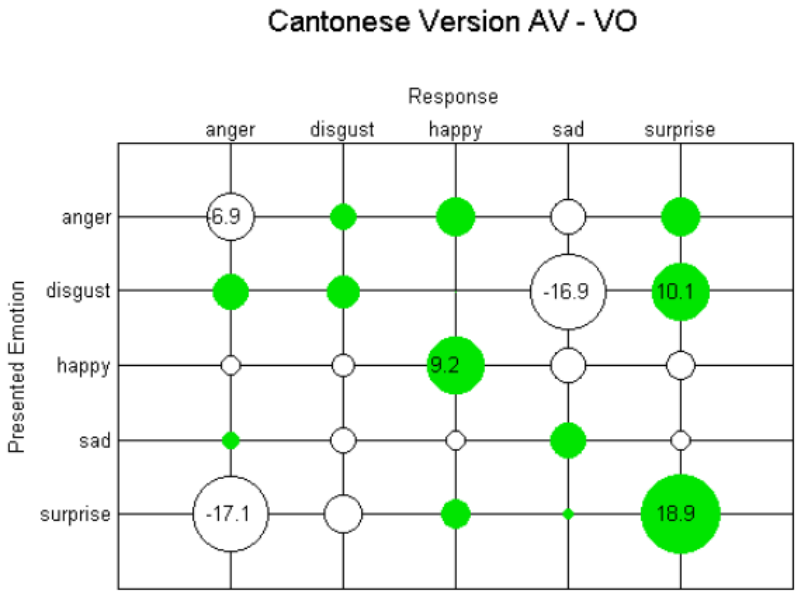

Figure 7. *Confusion matrix showing the differences in the percentage frequency of responses in the AV to VO conditions of Cantonese participants who participated in the Cantonese version of the experiment.*

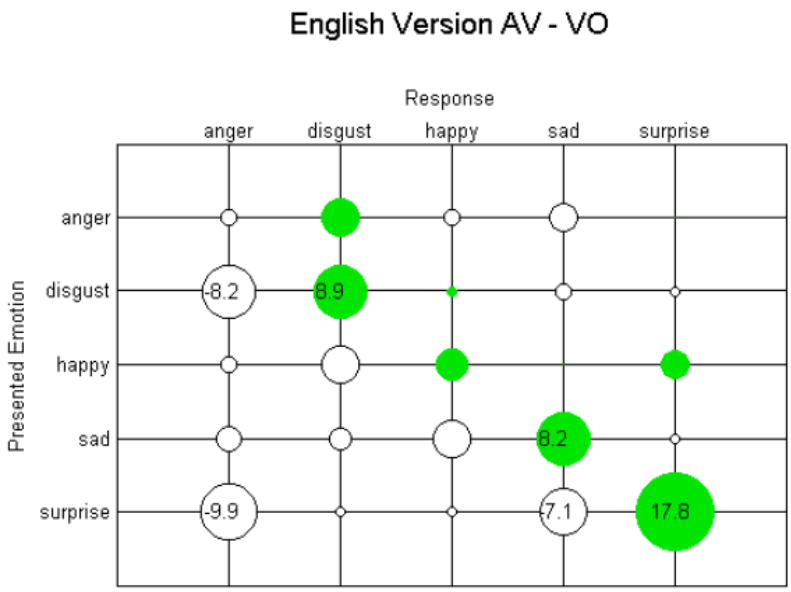

Figure 8. Confusion matrix showing the differences in the percentage frequency of responses in the AV to VO conditions of Cantonese participants who participated in the English version of the experiment.

Cantonese participants' overall accuracy in the English and Cantonese versions of the experiments. No significant differences between AV and VO were observed in both versions of the experiment F(1,19) = 1.55, *ns*, English F(1,19) = 1.92, *ns*.

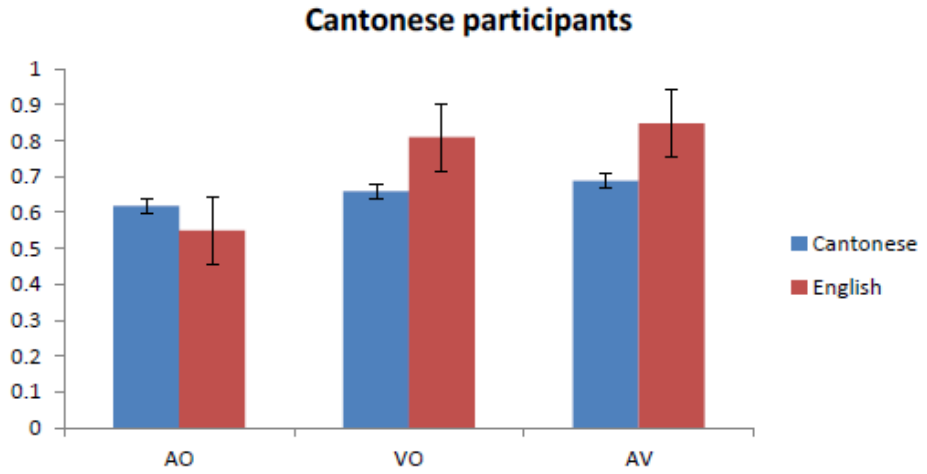

Figure 9. *Cantonese participants' accuracy in the English and Cantonese version of the experiment with standard error bars.*

Discussion

We found evidence that language plays a role in how we weigh up the importance of visual versus auditory emotion cues. Depending on the language an expression was expressed in, the Cantonese participants appeared to adopt two different strategies, an auditory bias effect when an emotion was expressed in their native language, but to give less weight to auditory cues when an emotion was presented in a non-native language. The same was the case for the Malay participants, but surprisingly this occurred in a direction opposite to what was predicted. That is, these participants were more likely to utilize auditory cues when judging the Cantonese expressions compared to English expressions. One explanation for this could be that, our sample of Malay participants from Malaysia was more exposed to Hong Kong accented Cantonese than Australian accented English since Chinese is the second largest ethnic population in Malaysia. It is also common for the media to broadcast material from

Hong Kong (movies, dramas and songs), so although none of the Malay participants speak Cantonese, they have a reasonable amount of exposure to Chinese culture and language.

It could be argued that the results of this study support the culture hypothesis since the Malay participants used more auditory cues when judging Cantonese expressions and the English participants showed a strong visual dominance effect, our results also indicate that the language hypothesis has some explanatory value. In this study, it appeared that the visual expressions in Cantonese speakers' productions were quite salient because the English speakers were able to pick out their facial expressions with a high degree of accuracy. It appears that the speakers were not abiding by the cultural display rules to suppress the expression of visual cues. So why then would the Malay participants in general and the Cantonese ones in particular, favour auditory information when it is likely less reliable than facial expressions? We propose that the auditory bias was simply because the Cantonese participants (and to a lesser extent the Malay ones) were able to understand either the semantics of the sentences or the emotion prosodic style of the language (or both). So rather than cultural display rules, it may be that familiarity with the language drives how these participants made use of the auditory information. On the other hand, the English participants showed a preference for facial expressions over vocal expressions regardless of language type. This visual dominance effect, whether it be a culture bias or not, is typical of the studies that have recruited participants from a Western culture [3, 1, 5, 6] and bears further examination.

Aside from the culture and language hypotheses, an interesting issue to explore is whether different emotion types are easier to recognize from the face or voice, thus driving the AV perception process. For example, consider the English participants' confusion matrix, when judging Cantonese expressions, it appears that auditory cues are key to disambiguating Disgust from Sad. The same can be observed in the Malay and Cantonese participants. This hints that different modalities may carry distinctive information that is emotion specific.

The question of whether the face and voice carry different information and how they may be integrated lies at the heart of our research interest. Our understanding of AV emotion perception is typically drawn from parallels with the AV speech perception literature. However this may be an inadequate model because speech perception and emotion perception may be very different things. Visual speech carries cues that are highly correlated with auditory information. For example lip gestures physically constrain the sounds that are produced; in contrast visual emotion cues are more independent of speech. Also, many visual emotion cues do not affect speech production per se, e.g. eyebrow movements, widening of eyes, etc. As such, visual and auditory emotion cues may carry the same emotional valence using different signals that are decoded separately. These different signals can carry unique information that are subject to its own patterns of confusion and when combined, produce an AV benefit. In this case, going back to our example above, although the Cantonese visual spoken expression of disgust can be confused as sad, the auditory expressions are not. So when given both visual and auditory information are present, the likelihood of the expression to be identified as sad is reduced, thus ruling it out as a potential competitor. So rather than a general culture or language bias, it may be worthwhile to explore if the weight assigned to visual and auditory cues may be emotion specific.

Conclusion

In conclusion, our results support the proposal that language can affect how information from different sensory modalities are weighed and used in making emotion judgments. Our results also suggest that there is interplay between language and culture that influence emotion perception.

References

[1] C. S. Chong, J. Kim, and C. Davis, "The effect of expression clarity and presentation modality on non-native vocal emotion perception," in *The 17th conference of the Oriental chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment /CASLRE (Conference on Asian Spoken Language Research and Evaluation)*. Phuket, Thailand: IEEE, 2014.

[2] B. De Gelder, K. B. E. Böcker, J. Tuomainen, M. Hensen, and J. Vroomen, "The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses," *Neuroscience Letters*, vol. 260, no. 2, pp. 133–136, 1999.

[3] J. Kim and C. Davis, "Perceiving emotion from a talker: How face and voice work together," *Visual Cognition*, vol. 20, no. 8, pp. 902–921, 2012.

[4] D. W. Massaro and P. B. Egan, "Perceiving affect from the voice and the face." *Psychonomic bulletin & review*, vol. 3, no. 2, pp. 215–21, Jun. 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24213870

[5] D. E. Bugental, J. W. Kaswan, and L. R. Love, "Perception of contradictory meanings conveyed by verbal and nonverbal channels." *Journal of Personality and Social Psychology*, vol. 16, no. 4, pp. 647–655, 1970.

[6] H. Ursula, K. Arvid, and S. Klaus R., *Multichannel communication of emotion: Synthetic signal production.* Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, 1988.

[7] O. Collignon, S. Girard, F. Gosselin, S. Roy, D. Saint-Amour, M. Lassonde, and F. Lepore, "Audio-visual integration of emotion expression." *Brain Research*, vol. 1242, pp. 126–35, Nov. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18495094

[8] A. Tanaka, A. Koizumi, H. Imai, S. Hiramatsu, E. Hiramoto, and B. de Gelder, "I feel your voice. Cultural differences in the multi- sensory perception of emotion." *Psychological Science: a journal of the American Psychological Society / APS*, vol. 21, no. 9, pp. 1259–1262, 2010.

[9] D. Matsumoto, S. H. Yoo, S. Hirayama, and G. Petrova, “Development and validation of a measure of display rule knowledge: the display rule assessment inventory.” *Emotion (Washington, D.C.)*, vol. 5, no. 1, pp. 23–40, 2005.

[10] M. D. Pell, L. Monetta, S. Paulmann and S. A. Kotz, “Recognizing emotions in a foreign language”,. Journal of Nonverbal Behavior, 33(2), 107-120, 2009.

[11] W. F. Thompson and L. L. Balkwill, “Decoding speech prosody in five languages.” Semiotica, 2006(158), 407-424, 2006.

[12] K. Scherer, R. Banse, and H. G. Wallbott, “Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures,” Journal of Cross-Cultural Psychology, vol. 32, no. 1, pp. 76–92, Jan. 2001.

[13] L. Anolli, F. Mantovani, and a. De Toni, “The Voice of Emotion in Chinese and Italian Young Adults,” *Journal of Cross-Cultural Psychology*, vol. 39, no. 5, pp. 565–598, Sep. 2008.

[14] E. D. Ross, A. E. Jerold, and G. Seibert, “The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice,” *Journal of Phonetics*, vol. 14, pp. 283–302, 1986.

[15] L., Anawin, “Intonation in Thai,” *D. Hirst and AD Cristo, Intonation Systems A Survey of Twenty Language*, pp. 376–394, 1998.

[16] T. Wang and Y. Lee, “Does restriction of pitch variation affect the perception of vocal emotions in Mandarin Chinese?” *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. EL117–EL123, 2015.

[17] “FFMPEG.” [Online]. Available: www.ffmpeg.org

[18] A. Lee, “Virtual Dub (Version 1.8. 6)[Software],” 2008.

[19] MATLAB 13.0, The MathWorks, Inc., Natick, Massachusetts, United States.

[20] C. Benoît, M. Grice, and V. Hazan, “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences,” *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996.

[21] L. L. N. Wong and S. D. Soli, “Development of the Cantonese Hearing In Noise Test (CHINT).” *Ear and Hearing*, vol. 26, no. 3, pp. 276–89, Jun. 2005.

[22] C. S. Chong, J. Kim, and C. Davis, “Development of an Audiovisual Cantonese Emotional Speech Database,” *The 17th conference of the Oriental chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment*

[23] K. I. Forster and J. C. Forster, “DMDX: A Windows display program with millisecond accuracy,” *Behavior Research Methods, Instruments, & Computers*, vol. 35, no. 1, pp. 116–124, 2003.

Perceptual similarity between participants

The general aim of the analyses conducted in this chapter was to determine if participants perceived emotion expressions in a more similar manner to those who spoke a similar type of language (i.e. non-tone language speakers, Malay and English) than to those who spoke a different type of language. In particular, we predicted that when identifying spoken expressions of emotions, the Malay speakers will be more likely to choose the same response options made by the English than the Cantonese participants.

While the analyses conducted in the previous section (accuracy and confusion matrices) are commonly used in emotion perception research, these methods however were inadequate in testing the prediction above due to a number of limitations. One, the measure of accuracy, when examined by itself, is a blunt measure that is insensitive to patterns in the distribution of response frequencies. For example, two groups of participants can have the exact same accuracy rates despite having very different response patterns, hence providing little information about the perception of emotions.

On the other hand, while examinations of confusion matrices may provide insights into how emotions are perceived; these are predominantly descriptive statistics which means that there is generally no consensus on how differences between confusion matrices may be quantified. The difficulty in quantifying differences further means that examination of confusion matrices may provide only indirect evidence of whether there were systematic differences in perceptions. Moreover, analyses of accuracy and confusion matrices are aggregate measures that do not provide any indication of whether differences in perception were driven by systematic or random differences in responses (i.e., perceptual similarity).

To address the above, I conducted an additional analysis on the data obtained from the study presented in the previous section. The aim of this analysis is to determine the extent to which

the responses of one group of participants match the responses of another group. In this analysis, agreement was measured using Fleiss' kappa which is a statistical measure that is typically used for assessing inter-rater reliability. For the purposes of the current work, inter-rater reliability is defined as a point estimate of the consensus or agreement in responses between groups of participants when identifying what they perceived to be the emotion expressed in each stimulus item of the experiment. Fleiss' kappa values were determined for each pairwise group of participants (i.e., Malay-Cantonese, Cantonese-English, and Malay-English) and a higher value is indicative of a greater degree of agreement in responses between the examined groups. The comparison of one kappa value to another provides an indication of whether there was a greater degree of agreement in responses in one pair of participant groups over another pair.

The predictions made for this analysis are similar to those made in the previous section. If there are systematic differences in how tone and non-tone language speakers perceive emotions, it is predicted that the kappa value for the agreement in responses between the non-tone language speakers, Malay and English (Malay-English) will be significantly greater than the kappa for the Malay-Cantonese participants. This pattern of results is expected in both Cantonese and English versions of the experiment and in both the AO and AV conditions.

Alternatively, if cultural display rules predispose those who share a similar culture to perceive emotions in a similar manner, the kappa value for the Malay-Cantonese participants will be significantly greater than the kappa for the Malay-English participants. However, given the results of the previous section, this is predicted to be the less likely alternative. Given the hypotheses, separate one-tailed t-test of Fleiss' Kappa values was conducted for each pairwise comparison group in each presentation condition and for each version of the experiment.

Results & Discussion

Table 1 shows the Fleiss' Kappa scores and the 95% confidence intervals for each pairwise group pairing for all presentation conditions of Cantonese spoken expression of emotions. In the VO condition, Fleiss' Kappa value did not differ significantly between the different groups (English-Malay vs. English-Cantonese, Z = -1.39, p = .08; English-Cantonese vs. Cantonese-Malay, Z = 0.01, p = .50; English-Malay vs. Cantonese-Malay, Z = -1.32, p = .09). This indicated that all three groups of participants generally perceived the facial expressions of Cantonese speakers in a similar manner.

Table 1. Response agreement as Fleiss's Kappa and 95% confidence intervals on Cantonese expressions of emotions.

		VO condition	
	Malay	**English**	**Cantonese**
Malay	-	0.54 (0.51 – 0.61)	0.49 (0.46 – 0.57)
Cantonese	0.49 (0.46 – 0.57)	0.49 (0.45 – 0.55)	-
		AO condition	
	Malay	**English**	**Cantonese**
Malay	-	0.46 (0.41 – 0.52)	0.38 (0.33 – 0.46)
Cantonese	0.38 (0.33 – 0.46)	0.39 (0.34 – 0.46)	-
		AV condition	
	Malay	**English**	**Cantonese**
Malay	-	0.68 (0.61 - 0.74)	0.36 (0.31 – 0.41)
Cantonese	0.36 (0.31 – 0.41)	0.37 (0.33 – 0.42)	-

In the AO condition, the Kappa value for the Malay-English speakers (*k* = 0.46, CI [0.41, 0.52]) was significantly higher than the Malay-Cantonese speakers (*k* = 0.38, CI [0.33, 0.46]) (Z = -1.84, p < .05). The difference between the Malay-Cantonese (0.38, CI [0.33, 0.46]) and the English-Cantonese groups (*k* = 0.39, CI [0.34, 0.46]) was not significant.

The pattern of results observed in the AV condition was similar to those in the AO condition. The agreement between the Malay-English speakers (*k* = 0.68, CI[0.61, 0.74]) was significantly higher than the agreement between the Malay-Cantonese speakers (*k* = 0.36 CI[0.31, 0.41]) (Z = -7.64, p < .001). No significant differences were observed in the comparisons between the Malay- Cantonese and the English-Cantonese groups, Z = .29, p = .39.

Table 2. Response agreement as Fleiss's Kappa and 95% confidence intervals on English expressions of emotions.

	VO condition		
	Malay	**English**	**Cantonese**
Malay	-	0.74 (0.68 – 0.79)	0.71 (0.64 – 0.77)
Cantonese	0.71 (0.64 – 0.77)	0.72 (0.67 – 0.77)	-
	AO condition		
	Malay	**English**	**Cantonese**
Malay	-	0.56 (0.50 – 0.62)	0.48 (0.42 – 0.54)
Cantonese	0.48 (0.42 – 0.54)	0.52 (0.46 – 0.59)	-
	AV condition		
	Malay	**English**	**Cantonese**
Malay	-	0.84 (0.80 - 0.89)	0.80 (0.75 – 0.84)
Cantonese	0.80 (0.75 – 0.84)	0.82 (0.78 – 0.86)	-

Table 2 shows the Fleiss' Kappa scores and the 95% confidence intervals for each pairwise group pairing for all presentation conditions of English spoken expression of emotions. Similar to the results of the Cantonese version of the experiment, Fleiss' Kappa values did not differ significantly between the different groups in the VO condition (Malay-Cantonese

vs. Malay-English, Z = 0.69, p = .25; Malay-Cantonese vs. Cantonese-English, Z = 0.24, p = .41; Cantonese-English vs. Malay-English, Z = -.53, p = .30).

In the AO condition, the kappa value for the Malay-English speakers (*k* = 0.56, CI[0.50, 0.62]) was significantly higher than the kappa for the Malay-Cantonese speakers (*k* = 0.48, CI[0.42, 0.54]) (Z = -1.85, p < .05). No significant differences were observed in the comparisons between the Malay-Cantonese and the English-Cantonese groups, Z = .89, p = .19. Somewhat different to the Cantonese version of the experiment, none of the Kappa values were significantly different In the AV condition (Malay-Cantonese vs. Malay-English, Z = 1.23, p = .11; Malay-Cantonese vs. Cantonese-English, Z = 0.65, p = .26; Cantonese-English vs. Malay-English, Z = -.65, p = .65). This may be due to the high accuracy rates (low variance in the distribution of response frequencies) of all participant groups in identifying AV English spoken expressions of emotions.

In summary, these results show that the perception of emotions was more similar between those who spoke a similar type of language (non-tone language, Malay and English participants) than those who spoke different types of languages (Malay and Cantonese participants). This difference were observed only when auditory information was perceivable (i.e., there was no differences in agreement in the VO condition). These results suggest that there are systematic differences in how tone and non-tone language speakers perceive emotions.

An important feature of these results is that it indicates, for the first time, that the differential use of F0 in the production of spoken emotions can lead to systematic differences in how tone and non-tone language users perceive spoken emotions. This finding is important as it helps us understand how differences in language specific characteristics may have an impact on spoken emotion recognition. One limitation of the current study is that it lacked a fourth

group of tone language speakers such as Mandarin speakers that would allow us to complement the findings of the study by determining if, similar to the non-tone language speakers, there is also a greater perceptual similarity among tone language speakers.

Chapter 3. Exploring Acoustic Differences between Cantonese (Tonal) and English (Non-Tonal) Spoken Expressions of Emotions

In the previous chapter, I demonstrated that there are systematic differences in how tone and non-tone language speakers perceive spoken expression of emotions. In this chapter I follow up on the results of the previous experiment by conducting an acoustic analysis study to examine how F0 is used in the production of neutral and emotion speech in Cantonese and English. In particular, I aimed to investigate the claim that in order to maintain the integrity of the properties of lexical tones, tone language users may use less F0 related cues when expressing emotions in speech (Ross, Edmondson & Seibert, 1986; Anawin, 1998; Wang & Lee, 2015[13]). To our knowledge few studies have directly compared F0 use in the production of spoken expressions of emotions in tone and non-tone languages, and none have examined Cantonese.

The second aim of the study reported in this chapter was to examine if a restriction in the use of F0 cues may lead to a strategically or systematically different way that they are used in the production of spoken expressions of emotions in Cantonese compared to English. That is, if the same emotion may be expressed using different F0 cues depending on the language of expression. Any systemic differences observed may explain the perceptual differences observed in the emotion recognition study presented in the previous chapter.

The work in this chapter was presented at Interspeech 2015, the 16th Annual Conference of the International Speech Communication Association held in Dresden, Germany. The paper

[13] These are listed in the reference list of the manuscript as [3], [4], [5].

was published in the peer-reviewed proceedings and is available in the ISCA archive[14]. This chapter reproduces the published proceeding with some minor edits.

[14] The proceeding is available at https://www.isca-speech.org/archive/interspeech_2015/i15_1522.html

Exploring Acoustic Differences between Cantonese (tone) and English (non-tone) Spoken Expressions of Emotions

Chee Seng Chong, Jeesun Kim, Chris Davis,
MARCS Institute,
University of Western Sydney

Abstract

It has been claimed that tone language speakers use less F0 related cues in the production of verbal expressions of emotions. This is because F0 is used in the production of lexical tones. This study investigated this claim by examining how F0 and various other acoustic parameters are used in the production of verbal emotion expressions in Cantonese (tone language) compared to English (non-tone language). Acoustic measurements (e.g., mean F0, F0 range) were extracted from the verbal expressions of five emotions (Angry, Happy, Sad, Surprise and Disgust) and a neutral expression produced by five male native speakers of Cantonese and English. Median F0 values and the number of peaks or troughs in the F0 contour per sentence were analysed using k-means clustering to determine how these properties are used in the production of spoken expressions of emotions and how they may vary as a function of language. The results showed some difference between the two languages in how F0 related cues are used in the production of emotions. The results are discussed in terms of the general acoustic characteristics of spoken emotion expressions and in relation to behavioural data from perceptual studies.

Introduction

Fundamental Frequency (F0) has been identified as one of the most salient carriers of vocal emotion information in the English language [1, 2]. However, this may not be the case for

tone languages because using F0 to express emotional prosody in expressive utterances may compromise the production of lexical tones [3, 4, 5]. This idea is supported by a study that found that the variance of F0 in emotion expressions and in a neutral baseline was much smaller in Mandarin (a tone language with 4 tones) than in Italian (non-tone language) [6].

The current study followed up this finding because a number of issues were raised by the Mandarin/Italian study. The first concerned the notion of whether F0 would really be over-loaded by having to code both emotion and lexical tone. In essence, it would depend on whether emotion and lexical tone were coded by the same type of variation in F0. In the Mandarin/Italian study [6], F0 variation was measured in terms of the range, minimum, mean, maximum and standard deviation of the F0. Of these measures, F0 range is quantified by the single point estimates of the minimum and maximum values and therefore insensitive to the total variation in F0. Moreover, the mean may not be a very good estimate because F0 ~~as it~~ is not normally distributed. In the current study, we compared acoustic measures of emotion expression in a tone and non-tone language and included two additional measures, median F0 and number of F0 turning points. The number of turning points is defined as the number of peaks or troughs in the F0 contour over a single sentence. As duration may vary across sentences, each utterance was divided into 30 equal time points and the F0 value was sampled at each point giving a measure of F0 every 80 to 100 ms.

The second issue concerns whether the finding that the expressions of emotions in Mandarin have a smaller F0 variation than Italian applies to the comparison between tone and non-tone languages in general. As the authors in [6] have pointed out, the Italian language has a particularly expressive speech style, one that is rich in F0 variation, so Italian may not be a typical representative of a non-tone language. Moreover, there is a dearth of studies on the expressions of emotions in tone languages, so determining whether this restriction in F0 generalizes to other tone languages is a worthwhile endeavour. Here then, we examined the

production of emotion in Cantonese, a tone language with 6 tones, with speakers of Australian English as the non-tone language comparison group. Cantonese is an interesting tone language to study as it has more lexical tones than Mandarin, so the linguistic system of this language may place greater demands on the use of F0.

Assuming that it is the case that all tone languages show restricted F0 variation in emotion expression, the third issue then becomes what is the best explanation for this? A recurring claim is that this restriction preserves the signal integrity of the lexical tones. However to our knowledge, this claim has never been verified, so we tested this idea by examining the differences in F0 variability of neutral expressions between tone and non-tone languages. The basic premise that underlies the emotion F0 restriction argument is that neutral expressions in Cantonese should have a larger F0 range than English, i.e., that the linguistic system has monopolized the use of F0. In other words, due to the constraints imposed by the linguistic system, there is a diminished opportunity for F0 to carry emotion information, thus leading to a restricted use of F0 in emotion expression.

The final issue concerns the role of F0 in emotion expression. Despite the finding that there is a difference in F0 variation in tone language expressions of emotions, no clear evidence was provided concerning the role that F0 plays in the production of emotions. As the authors of [6] pointed out, it may be although F0 is still used in Mandarin to some extent, other cues such as intensity and speech rate may be used to supplement F0. Alternatively, due to the restriction in F0, tone language users may utilize F0 cues in a manner different from non-tone language users. For example, it has been shown that Cantonese speakers raise their mean F0 to convey sarcasm while English speakers tend to lower their mean F0 [7]. So there is a possibility that the same F0 cues may be used to express different emotions depending on the language. To investigate the mix of acoustic factors used in emotion expression, we used K-means cluster analysis to examine at how the different acoustic properties are grouped

together and whether there is a difference in how F0 related properties are used across the languages.

Methods

Participants

Cantonese participants: Five male native speakers of Cantonese who were born and raised in Hong Kong were invited to participate for monetary reimbursement[15]. The average age of the participants was 29.1 years.

English participants: Emotion expressions of five male native speakers of Australian English were obtained from our lab database. The average age of the participants was 23.0 years.

Materials

Speech Materials

Fifty semantically neutral sentences were chosen from the Cantonese Hearing In Noise (CHINT) sentences list [8] on the basis that they had a good spread of different tones at the initial and final position in the sentences (see for detailed procedure). All 50 sentences were recorded as expressive speech stimuli for the five basic emotions and neutral sentences.

Ten sentences selected from the Semantically Unpredictable Sentences [9] were recorded as the English expressive speech stimuli for five emotions (Anger, Disgust, Happy, Sad, Surprise) and the Neutral expression.

Production Setup

While only the audio recordings are used in this paper, the entire recording procedure including video recording is reported for completeness. Participants were seated in front of a 20.1" LCD video monitor (Diamond Digital DV201B) that is used to present the stimulus

[15] Extracted from the CAVES database

sentences to the participant. Directly above the monitor was a video camera (Sony NXCAM HXR-NX30p) where participants were requested to fixate at prior to expressing the sentences. The videos were recorded at 1920 x 1080 full HD resolution at 50 fps. To capture participants utterances a microphone (AT 4033a Transformerless Capacitor Studio Microphone) was placed about 20 cm away from the participants' lips and out of the field of view of the camera. Audio captured using the microphone was fed into the Motu Ultralite mk3 audio interface with FireWire connection to a PC running CueMix FX digital mixer and then to Audacity which captured the sound at a sampling rate of 48kHZ. This audio feed as well as video feed from the video camera was monitored by the experimenter outside of the booth who provided the participants with feedback as well as displaying the next sentence on the monitor in front of the participants. The recording details of English speakers were similar to the Cantonese recordings (see [10] for details).

Procedure

Production of emotions

Since verbal expressions of emotions are often consciously and deliberately produced to convey emotions, rather than focusing on the experience of the emotion itself, we were interested in emotional expression; the signals that people present to others to express emotion. Given this, participants were instructed to be as natural as possible in how they expressed themselves and were asked to produce the emotions with the intent of communicating their emotional feelings to an observer. Moreover, through the course of the recording, the experimenter did not interfere, guide or give examples as to how each emotion type should be expressed to avoid demand characteristics and to preserve the natural idiosyncratic variation in the expression of emotions.

During the recording sessions, each stimulus sentence was displayed one at a time in a random order on the computer monitor and the participants then produced the utterances

when ready. Participants were given feedback via the screen if they had to repeat the sentence (e.g., they misread the sentence or did not fixate on the camera while producing the expressions). Participants were also given a short story as a form of mood induction and three practice trials prior to the start of each emotion block and asked to put themselves in the mode of expressing the emotion. As was mentioned, the emotions to be expressed were blocked, so participants would produce all 50 sentences expressing the same emotion giving a total of 350 sentences per speaker (50 sentences x 7 (six emotions plus neutral).

Analysis

Acoustic parameters (mean F0, minimum F0, maximum F0, duration, maximum velocity, and mean intensity) of whole sentences were automatically extracted using Prosody Pro [11], a Praat script [12] with the lower threshold of F0 set at 50 Hz and the upper at 350 Hz. Preliminary checks found that the range of F0 values for all of the examined stimuli fell within the 50 to 350 Hz range. We then extracted the median, number of turning points and speech rate (duration/ number of syllables) using a Matlab [13] script.

Results

Do Cantonese neutral expressions differ from English in F0 related measures?

Using a multivariate analysis of variance, Pillai's trace showed a significant effect of language on the mean, minimum, maximum, median and number of turning points of F0, V = 0.91, $F(4,5) = 7.72$, $p < 0.05$. However, separate univariate ANOVAs revealed non-significant effects of languages on each of the F0 measures, minimum, $F(1,8) = 0.61$, *ns*, maximum, $F(1,8) = 0.02$, *ns*, mean, $F(1,8) = 0.03$, *ns*, median, $F(1,8) = 0.01$, *ns*, mean, $F(1,8) = 0.03$, *ns*, number of turning points $F(1,8) = 0.92$, *ns*. Table 1 below shows the mean values for all of the measures.

Table 1: *The mean values of minimum, mean, maximum, median (in Hz) and number of turning points of neutral expressions.*

Language	F0 mean	F0 min	F0 max	F0 median	T.points
English	132.68	96.71	182.62	127.04	10.94
Cantonese	130.75	93.18	180.55	128.14	11.52

The MANOVA was followed up with a discriminant analysis which revealed that the language groups can be differentiated by mean F0 (b=1.47) and median F0 (b=-1.58), i.e., although there was no difference in F0 range, English neutral expressions had higher mean F0 but lower median F0 than the Cantonese expressions.

Do Cantonese emotion expressions use a smaller range of F0 derived cues?

Separate MANOVAs revealed that except for happy, V = 0.84, $F(4,5) = 4.26$, *ns*., and surprise, V = 0.90, $F(4,5) = 7.22$, *ns*, there was a significant difference between Cantonese and English on the F0 measures in expressions of angry, V = 0.97, $F(4,5) = 24.3$, $p < 0.01$, disgust, V = 0.96, $F(4,5) = 17.47$, $p < 0.01$, and sad, V = 0.98, $F(4,5) = 33.57$, $p < 0.01$. All significant effects were followed up with ANOVAs. Table 2 lists the results of the ANOVAs.

Both Angry and Sad English expressions had a significantly higher number of F0 turning points than did the Cantonese expressions, (14.24 vs. 11.23 and 12.66 vs. 10.35, respectively). For Disgust, the median F0 value was higher for the English expressions (168.52 Hz vs. 120.38 Hz).

Table 2: *F values and significance levels for the difference between English and Cantonese F0 values for each emotion expression.*

Emotion	min	max	mean	median	T.points
Angry	.16	.37	.10	.22	27.24 ***
Disgust	.83	4.3	.01	21.25 **	.42
Happy	*ns*	*ns*	*ns*	*ns*	*ns*
Sad	.69	.02	.25	.45	41.8 ***
Surprise	*ns*	*ns*	*ns*	*ns*	*ns*

Note: The values in the table are for F(1,8), *p < .05, **p < .01 ***p < .001.

In summary, with regards to F0, the difference between English and Cantonese neutral expressions is captured by median F0 and mean F0, indicating that the difference between the two lie in the contour shapes of F0. The measures of median F0 and F0 turning points were able to capture the difference between English and Cantonese expressions of Angry, Disgust and Sad. When compared with the F0 measures of the neutral expressions, it is clear that the median F0 and number of turning points in Cantonese emotion expressions did not vary much from the neutral expressions. On the contrary, English emotion expressions had a larger median F0 and higher number of turning points than the neutral expressions.

k-means clustering

The data consisted of 10 acoustic measures, (duration, speech rate, final F0, minimum F0, maximum F0, max F0 velocity, mean F0, mean intensity, median F0 and number of F0 turning points), extracted from the 10 speakers by 5 emotion and 1 neutral expression giving a total of 1800 utterances. Separate k-means analyses were conducted for each language.

On the first pass using k=6 (emotions), instead of emotion type, the speakers themselves were identified as the most salient clusters, accounting for 86.3% (Cantonese) and 79.4% (English) of the variance. So we conducted separate k-means analysis for each of the speakers. Given the scope of the current study and that the only F0 measures that showed a clear difference between Cantonese and English expressions were the median and number of turning points; we focused the analysis on these two factors. Here we examined how these measures were clustered, specifically, we examined which emotion was classified to have the highest and lowest centroid means for these measures and if this differed between Cantonese and English. The cluster solution for each speaker is listed in Table 3.

Table 3: *Centroid means for the cluster solution for each speaker (median F0 and Turning point data).*

Speaker	Median		T.points	
	highest	lowest	highest	lowest
Eng1	angry	happy	happy	neutral
Eng2	angry	happy	happy	neutral
Eng3	angry	happy	happy	neutral
Eng4	surprise	neutral	angry	surprise
Eng5	sad	neutral	surprise	neutral
Cant1	happy	neutral	neutral	angry
Cant2	angry	neutral	neutral	sad
Cant3	happy	disgust	disgust	sad
Cant4	sad	neutral	disgust	surprise
Cant5	sad	surprise	disgust	angry

From Table 3, it is clear that 3 out of the 5 English speakers in the study produced Angry, Happy and Neutral quite similarly (Eng1, 2, and 3). Although there were only 5 speakers per group, the Cantonese speakers provided a stark contrast. First, among the Cantonese speakers, there was little agreement as to what emotion corresponded to a high median (2 happy, 2 sad, 1 angry) or low number of turning points (2 sad, 2 angry, 1 surprise). Second, other than Neutral having the lowest median value, there was little agreement

between English and Cantonese speakers. It is interesting to note that there was some similarity between Cantonese and English expressions such that, an emotion that had the lowest median F0 also tended to have a highest number of F0 turning points. This suggests that although expressions of emotions in Cantonese and English may utilize similar clusters of acoustic cues, they do not convey the same emotion.

Discussion

In this study, we aimed to investigate the differences between a tone and a non-tone language in terms of the differences in the manner that F0 was employed in speech. The motivation for this study was to follow up a study that had examined differences between Mandarin and Italian expressions of emotions by addressing four specific issues. First of all, we were concerned with how F0 variance may best be captured. Instead of relying on single point estimates such as range and mean F0, we included two additional measures that we judged would better capture the relevant aspect of variance, namely, the median F0 and number of F0 turning points. The median is a better estimate of the F0 because F0 is not normally distributed. Moreover it is more meaningful to define F0 in terms of fluctuations or contour change over time. Therefore, an estimate of the number of turning points allowed us to capture this change over time at the sentential level. Indeed throughout the analysis conducted in this study, we found the median F0 scores and number of F0 turning points to be the F0 derived features that best discriminated emotions and the language in which these were expressed. On this point we propose that it is important for future studies on verbal expressions of emotions to use more sophisticated measures that take the contour of the F0 and its fluctuations over time into consideration.

Concerning the second issue of whether a restriction of F0 variation can be generalized to other tone and non-tone languages, we found partial support. This restriction was not

observed across all emotions and what restriction we found was only for the median F0 and number of turns. Whether this constitutes as a restriction of variation is debatable because no differences were observed on the other measures such as range and mean. Our results do however suggest that the F0 contour may be different, i.e., Cantonese expressions of emotions may be flatter or more monotonous than English emotion expressions. The third issue concerned how the linguistic properties of Cantonese as a tone language may affect F0 use in emotion expression. We examined and found no evidence that Cantonese use more F0 related cues resulting in a diminished capacity for F0 to carry emotion information. This finding is interesting as it is goes against the idea that tone languages may have a larger range or a larger number of turning points given the nature of its linguistic property. Discriminant analysis however suggests that a combination of F0 measures (mean and median) can distinguish Cantonese from English neutral expressions. This once again emphases the point that single measures of F0 may be uninformative, as it is the variation of F0 over time that appear to be more important.

Given that in terms of the manner in which F0 was employed in Cantonese neutral expressions were similar to that in English, it follows that Cantonese should have the freedom to use F0 for emotion prosody in a manner similar to English. Yet, the data showed that instead of using the F0 space (or at least as much as English emotion expressions do); Cantonese expressions of emotions used less F0 variation. This seems likely due to the need for Cantonese speakers to modulate their use of F0 to preserve the signal properties of lexical tones, although here a more sophisticated measure of the mix and interaction of tonal acoustic properties are needed.

Finally with regards to the role that F0 plays, we examined how the clusters of acoustic properties may differ in Cantonese and English expressions of emotions. In this study we only examined median F0 and number of F0 turning points. The solutions for our k-means cluster

analysis showed that these measures were similarly grouped in Cantonese and English expressions of emotions. However, interestingly these clusters corresponded to different emotions. This was particularly the case if we considered expressions that had low median F0 (Neutral for Cantonese and Happy for English). As for high median F0, the clusters corresponded to the expression of Anger in English but did not fit any emotion in the Cantonese expressions. This is also interesting because this finding fits the previous observation that Cantonese speakers may use less F0 variation. So instead of coding very active expressions like Angry that is usually associated with high F0, it may be that Cantonese uses F0 mainly for conveying less active emotions like Happy or Neutral, and may use a more moderate or mid-range F0 that can convey emotion without distorting the boundaries of lexical tones. Whereas for emotions like Anger or Sadness that is usually coded by the maximum or minimum F0, features such as speech rate and intensity may be used as the most salient feature instead.

In the literature, it is a common finding that people are poorer at recognizing emotions produced in a language that one is not familiar with [14, 15]. This effect often produces systematic perceptual confusion between emotions. We propose that part of this confusion may be due to linguistic effects on emotion production. By combining the results of production and perception studies, we aim to provide a more unified understanding of how linguistic properties affect the production of emotion expressions.

Conclusion

In conclusion, our results gave some support for the proposal that emotion expression in tone languages has less F0 variation than in non-tone languages. However, the picture that emerged of the way that F0 was used for the expression of emotion in tone vs. non-tone languages was rather complex.

References

[1] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: different channels, same code?" *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814, Sep. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12956543

[2] K. R. Scherer and J. S. Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," *Motivation and Emotion*, vol. 1, no. 4, pp. 331–346, Dec. 1977. [Online]. Available: http://link.springer.com/10.1007/BF00992539

[3] E. D. Ross, A. E. Jerold, and G. Seibert, "The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice," *Journal of Phonetics*, vol. 14, pp. 283–302, 1986.

[4] L., Anawin, "Intonation in Thai," *D. Hirst and AD Cristo, Intonation Systems A Survey of Twenty Language*, pp. 376–394, 1998.

[5] T. Wang and Y.-c. Lee, "Does restriction of pitch variation affect the perception of vocal emotions in Mandarin Chinese?" *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. EL117–EL123, 2015.

[6] L. Anolli, F. Mantovani, and A. De Toni, "The Voice of Emotion in Chinese and Italian Young Adults," *Journal of Cross-Cultural Psychology*, vol. 39, no. 5, pp. 565–598, Sep. 2008. [Online]. Available: http://jcc.sagepub.com/cgi/doi/10.1177/0022022108321178

[7] H. S. Cheang and M. D. Pell, "Acoustic markers of sarcasm in Cantonese and English." *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1394–405, Sep. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19739753

[8] L. L. N. Wong and S. D. Soli, "Development of the Cantonese Hearing In Noise Test (CHINT)." *Ear and hearing*, vol. 26, no. 3, pp. 276–89, Jun. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15937409

[9] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996.

[10] J. Kim and C. Davis, “Perceiving emotion from a talker: How face and voice work together,” *Visual Cognition*, vol. 20, no. 8, pp. 902–921, 2012.

[11] I. Xu, “ProsodyPro A Tool for Large-scale Systematic Prosody Analysis,” in *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, Aix-en-Provence, France, 2013, pp. 7–10.

[12] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program],” 2014. [Online]. Available: http://www.praat.org/

[13] MATLAB 13.0, The MathWorks, Inc., Natick, Massachusetts, United States.

[14] C. S. Chong, J. Kim, and C. Davis, “The effect of expression clarity and presentation modality on non-native vocal emotion perception,” in *The 17th conference of the Oriental chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment /CASLRE (Conference on Asian Spoken Language Research and Evaluation)*. Phuket, Thailand: IEEE, 2014.

[15] K. Scherer, R. Banse, and H. Wallbott, “Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures,” *Journal of Cross-Cultural Psychology*, vol. 32, no. 1, pp.76–92, 2001.

Chapter 4. Disgust Expressive Speech: The Acoustic Consequences of the Facial Expression of Emotion

An overarching theme of this thesis is to examine how the perception and production of spoken expressions of emotions may be shaped by the conjoint, and often competing, demands of emotion expression and speech production. In the previous chapters, I have focussed on one language specific feature, examining how the acoustic property of F0 may be used differently in the production of expressive speech in tone and non-tone languages. In this chapter, I examined another way in which expressive speech may be affected by the competing demands of emotion expression and speech production.

While the previous chapters focussed on the auditory domain (i.e., language), this chapter takes a slightly different approach, examining the role that facial expressions of emotions play in shaping the acoustics of expressive speech. Here I tested the general idea that in expressive speech, the production of facial expressions may interfere with how we shape our lips to articulate speech sounds. In particular, this chapter examined how the emblematic facial expression of disgust may affect the configuration of the lips during speech articulation and how this may in turn affect the first and second formant frequencies (F1 and F2) of speech.

I chose to examine disgust because the prototypical facial expression of disgust heavily involves the lower half of the face and in particular, our lips are thought to act as a physical barrier that prevents the entry of contaminants into the mouth cavity. Moreover, as it is argued that this facial expression serves a functional purpose with respect to the preservation of our wellbeing, the selection pressure for such gestures may take precedence over aspects of speech articulation. Disgust would therefore likely be one of the emotion types that impose the clearest and largest effects on lip articulation.

The work presented in this chapter was published in Speech Communication in 2018[16] and the article is reproduced in full with some minor edits. This work was also developed as an extension of a study that was presented at Interspeech 2016, the 17th Annual Conference of the International Speech Communication Association held in San Francisco, USA[17].

[16] The full journal article is available at https://www.sciencedirect.com/science/article/abs/pii /S0167639317300420

[17] The proceeding is available at https://www.isca-speech.org/archive/Interspeech_2016/pdfs/1463.PDF

Disgust Expressive Speech: The Acoustic Consequences of the Facial Expression of Emotion

Chee Seng Chong, Jeesun Kim, Chris Davis,
MARCS Institute,
Western Sydney University

Abstract

This study investigated how the facial expression of disgust may affect the acoustics of speech. In terms of a pathogen avoidance mechanism, the expression of disgust would seem to require speech to be produced with a smaller mouth opening than neutral speech, hence lowering the formant frequencies. This hypothesis was tested by comparing how lip configuration (i.e., height, width and size of the lip area), fundamental frequency (F0) and the formants (F1 and F2) of the vowels ([ɐ], [ɛ:], [i:], [ɔ:], [u:]) changed when produced in neutral or disgust expressions. The vowels were extracted from 50 Cantonese sentences spoken by 10 (5 male) talkers; produced once in disgust and once more in a neutral tone of voice. The results support the notion that the facial expression of emotions may have a role in shaping the acoustic properties of the vocal expressions of emotions. Mixed effects logistic regression models revealed that in disgust, vowels were produced with lower lip height, lower F1, F2, and higher F0 than neutral speech.

Introduction

This study aims to understand the properties of auditory and visual expressive speech that result from the simultaneous expression of speech and emotion. Our focus is on disgust, an emotion that is expressed through facial features that typically involves marked changes in mouth area and thus is likely to interact with speech articulation (affecting the properties of

auditory and visual speech). Our approach was to first quantify the lip and mouth movements of disgust expressive speech and then examine how such actions may modify the acoustics of speech.

Disgust is an emotion type that is claimed to have a clear evolutionary underpinning. It is widely held that disgust evolved as a pathogen avoidance mechanism [1, 2] and thus the types of stimuli that elicit disgust [3] and the way it is expressed is similar across cultures [4]. The expression of disgust involves the wrinkling of the nose, closure of the lips to prevent access to the vocal cavity, and tongue extrusion to facilitate expulsion of foreign agents from the body. This has led to claims that the emblematic facial expression of disgust serves the goal of reducing the probability that a contaminant or pathogen may enter our bodies [5, 6, 7].

It is clear that the facial expression of disgust involves the lower half of the face and that the lips appear to act as a physical barrier that prevents the entry of contaminants. Hence, the concurrent production of speech and the disgust emotion expression is likely to impose competing demands on lip articulation. That is, while disgust requires a specific and constrained lip configuration in order to reduce the mouth aperture, the production of speech sounds requires the lips to dynamically assume different configurations. Precisely how these expressive and articulatory demands are resolved is not well attested. There is however, some evidence to suggest that the expression of disgust can affect the configuration of the articulators. For example, a facial motion capture study revealed that compared to neutral speech, spoken expressions of disgust are generally produced with significantly more advanced jaw, nose wrinkling, upper and corner lip raising, and lowering of the larynx [8]. Looking more specifically at lip configurations using reflective markers, a study by Caldognetto and colleagues [9] found that vowels are produced with minimal mouth opening, maximal spreading and retraction of the lips, and negative right vertical asymmetry when produced in disgust.

Given that the facial expression of disgust serves a functional purpose with respect to the preservation of our wellbeing, our hypothesis is that the selection pressure for such gestures may take precedence over aspects of speech articulation. That is, disgust may impose a restriction on the configuration of the lips during articulation, such that the production of speech sounds, especially those that require mouth opening, i.e., vowel sounds, may be compromised. Considering that there is an explicit link between the size of the mouth opening and the frequency of the first formant (F1) [10], the effect of any restriction in spoken lip configuration will most likely result in measurable changes in the formant values.

The idea that facial expressions may affect speech formant values is not a novel one. For example, it has been claimed that speech produced while smiling has higher pitch (F0) and formant frequencies compared to neutral speech. It is claimed that smiling shortens the vocal tracts thereby resulting in a change in formant frequencies [11]. An analogous effect was observed in a later study; utterances produced while frowning (smaller mouth opening) were observed to have lower formant values than neutral utterances [12]. Surprisingly, despite the fact that the expression of disgust clearly affects the mouth region, few studies have explored if the vocal expressions disgust is associated with changes in formant frequencies. For instance, in a review of 104 studies [13], only one study had examined this possibility [14]. In this study, acoustic analysis of the vowels [a], [o], and [e] produced by four speakers showed that the vowel [e] in disgust had higher formant values (second format frequency, F2), whereas the vowel [o] had a lower F1. The study also reported that disgust was best recognised in the vowel [o]. These results are consistent with what we have reported in a preliminary study which compared disgust and neutral speech of five female speakers [15]. In this study, we found the vowel [ɔ:] to have the largest changes in formant frequencies. We suggested that open and rounded vowels such as [ɔ:] are produced through a configuration

that conflicts with what is conditioned by the expression of disgust, hence showing the largest formant changes.

The current study is a follow up of our earlier one [15]. Here we not only examined auditory but also the visual properties of spoken expressions of disgust. The examination of visual properties is important particularly since none of the above studies have examined if spoken expressions of emotions actually involved changes in mouth opening when compared to neutral speech. In the study on smiled speech [11], the assumption between smiling and increased mouth opening during speech production was made based on a study of non-verbal smiles by Shor [16]. Likewise, the relationship between lip configuration and the expression of disgust was an untested assumption in [15]. In the current study, we analysed the auditory and visual parameters associated with disgust at the middle of sustained vowels. We first conducted a visual analysis of the face to verify if disgust expressive speech was indeed produced with smaller aperture and/or lip height. We then examined if the predicted lowering of formants may be observed.

Extending the preliminary study [15], the current study included data from an additional five male speakers (a total 10 speakers). F0, F1 and F2 were measured from five vowels ([ɐ], [ɛ:], [i:], [ɔ:], [u:]) extracted from 50 sentences (produced once in a disgust and once in a neutral tone of voice). Visual analysis of the lips and mouth opening was conducted by measuring the size, height and width of a region enclosing the lips during the production of those five vowels. Our prediction was that disgust expression would involve a contraction around the mouth region such that speech is produced with a smaller mouth opening and/or with a reduction in lip height; and that auditory speech will have lower formant frequencies (especially F1), when compared to neutral speech.

Methods

Material

Speech materials were obtained from the Cantonese Audio-Visual Expressive (CAVE) speech database [17] which contains audio-visual recordings of 10 native speakers of Cantonese (five females, mean age = 29.1, SD = 4.9, none of the speakers reported histories of speech, language, or hearing problems) expressing 50 semantically neutral Cantonese sentences in different tones of voices (six basic emotions and neutral). The sentences were selected from the Cantonese Hearing In Noise Test sentences list [18] on the basis that they have a good distribution of tones in each sentence. For the purposes of this study, only disgust and neutral utterances of ten speakers were selected, giving a total of 1000 audio-visual spoken sentences (10 speakers x 50 sentences x 2 emotions (disgust, neutral)).

In developing this database, the speakers were encouraged to express themselves as naturally as possible with the intent of communicating their emotions to an observer. They expressed each sentence at their own pace and were given the opportunity to repeat an utterance until they were satisfied that their portrayal was similar to how they would have expressed it outside of an experimental setting.

Acoustic Measure

This analysis is similar to what was reported in [15]. The Cantonese sentences were first transcribed into Jyutping (a romanisation system for Cantonese[18]) and then altered to the closest approximation of Spanish SAMPA.

The transcriptions were then force-aligned using EasyAlign and manually checked and corrected by the first author. 76 utterances were removed due to mispronunciation or missing

[18] http://www.lshk.org/jyutping

data (a word omitted in the recording). This yielded a final data count of 1000 instances of [ɐ], 382 of [ɛː], 1398 of [iː], 896 of [ɔː] and 406 of [uː], half of these were produced in disgust and the other with neutral expression. By comparing a vowel produced in disgust with the same vowel produced in neutral within the same context word and sentence would reduce any potential coarticulation effects. Mean F0 values for the sustained vowels were extracted using ProsodyPro [21], a script implemented in Praat. Mean formant values were extracted using Linear Predictive Coding (LPC) analysis implemented through Burgs algorithm in Praat.

Visual Measure

The goal of this analysis was to quantify changes in the mouth region. A single video frame was extracted at the mid-point of a sustained vowel and three measurements of the lips were made. The size or area of the mouth region was defined by the number of pixels occupied by the lips and the enclosed mouth opening. A bounding box was then applied to the region; with the lip height corresponding to the height of the bounding box and lip width to the width of the box. Figure 1 below shows an example of how the measurements were taken. The area of the lips is given by the number of pixels occupied by the enclosed free-form red boundary around the lips, while lip height and width is defined as the height and width of the green bounding box around the lip region.

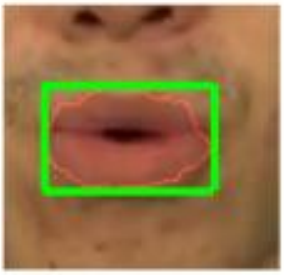

Figure 1. Sample output showing how the lip region was measured.

Due to the amount of data and to minimize human error, the above steps were automated using a Matlab script[19]. The image frames of interest were extracted using FFMPEG based on the timing information obtained from the aligned Praat textgrids. Texture segmentation, edge detection and a RGB separation algorithm were applied to automatically detect and segment the lips (using a custom Matlab script). The segmented regions were saved as a separate image file for visual inspection (see Figure 1).

Since the lip measurements were made on a 2D approximation of a 3D object, it may be subject to error due to the speaker's head pose, distance from the camera, and so on. In other words, the measurement may be noisy due to variations in speakers' distance from the camera or camera zoom. In order to reduce the impact of these factors, visual screening was manually performed by one of the authors to select only images where speakers had a forward facing head orientation with both ears equally visible, i.e., no obvious rotation in the pitch, yaw and roll axes. This reduced the data set to 487 images of [ɐ], 262 of [ɛː], 766 of [iː], 857 of [ɔː] and 403 of [uː].

In order to compensate for any changes in the distance of the speaker from the camera, a normalisation procedure was adopted. The aim was to ensure that the area of the face region for each speaker across emotion conditions would be normalised so that the size of the mouth region could be appropriately compared. This was done by randomly selecting three images of a speaker producing a vowel type in a neutral tone of voice condition. The area, height and width of the face region were measured manually and the average of each measure determined. This procedure was repeated for the same speaker producing the same vowel type for the disgust condition. The difference between the means was then expressed as a ratio. The ratio differences were then determined for the rest of the vowels and speakers. A

[19] The algorithm is described in the next chapter.

small ratio means that the size of the speaker's face, and by proxy, the mouth region was relatively consistent across the two emotion types or recording sessions. This would also suggest a negligible change in the speakers distance from the camera. Across all speakers and vowels, the differences between recordings were small, an average of 3.69% difference in area, 3.37% in height and 4.60% in length. To adjust for these small differences, all lip measurements were corrected by multiplying the appropriate ratios by the raw measurements.

Note that our measures of lip area include both the aperture of the mouth opening and surrounding lips. No difference in these measures should be observed if the expression of disgust does not affect articulation. On the contrary, if any difference was observed, it should be attributable to mouth opening as the size of a person's lips should be constant when articulating the same vowel across recordings. Moreover, the measures of lip height and width should provide a more nuanced estimate of the lip configuration and mouth aperture.

Data Cleaning

To identify speaker and vowel specific outliers, we examined each of the five vowels separately by emotion type and speaker. Data points that were larger or smaller than 1.5 times the interquartile range of the set were considered outliers and were removed from analysis. A total of 532 and 274 items were removed from the acoustic and visual data set, respectively. This resulted in a final data set of 3550 acoustic and 2501 items visual data points. In order to ensure that this study has a power of at least 0.8, the visual and acoustic analysis used a separate subset of items as only a moderate number of items were available in both modalities after the cleaning procedures.

Analysis

Acoustic

Measures of speech frequency tend to be subject to large idiosyncratic individual differences and these were removed by mean-centering the data using the relevant speaker and vowel

specific mean values. For example, each measure of F0 for each token utterance of [ɐ] produced by speaker 1 in both neutral and disgust was mean centered using the mean F0 value for all utterances of [ɐ] produced by speaker 1 in the neutral condition. This was performed for all acoustic measures (F0, F1 and F2) and for all speakers. The null hypothesis predicts that there will be a significant difference between the neutral and disgust conditions where the acoustic measures for disgust will show a larger negative change (reduction in frequency) compared to neutral expressions.

Visual

It is important to note is that the measures of lip configuration used in this study was made in pixels, a unit that is hard to interpret, i.e., a 20 pixel difference is not meaningful as it is relative to the individual, distance from camera, and resolution of the video clip. So the lip measures for each speaker producing each vowel were converted into Z-scores. This allowed us to examine if changes in the lip configuration is a salient predictor of whether an utterance was produced in disgust or neutral. The conversion to Z-scores eliminates idiosyncratic individual differences and facilitates model convergence and interpretation of the fit.

The null hypothesis is that regardless of the emotion expressed, utterances will be equally distributed across the entire Z-distribution; hence the various measures will not be predictive of emotion type. On the other hand, the experimental hypothesis predicts that utterances will be clustered by emotion type where the likelihood that an utterance was produced in disgust increases as a measure approaches the lower end of the Z-distribution. In other words, the smaller the area, width or height measured, the more likely it is to be a disgust utterance.

Results

For the acoustic analysis, a MANOVA was conducted for each vowel type with emotion, lexical tone (there are 6 lexical tones) and gender entered as independent variables. For the

visual analysis, a mixed effects logistic regression model was fitted to each vowel type using the lme4 [22] package in R. A backward selection procedure was used to determine the model that best fit the data. An initial model that includes all factors, fixed (Z-transformed area, height and width) and random effects (gender and speaker) with a binary dependent variable of Disgust or Neutral were entered into the model. Subsequent models were fitted by removing the variable with the largest p-value. The simpler model was retained if the fit of the model did not change significantly. These steps were repeated until the most parsimonious model was obtained.

The results of these analyses are reported in Table 1 below. Gender specific results were provided where an interaction between gender and emotion was found. β estimates, log odds ratios and Z values were reported for the mixed effects logistic regression models while the difference in Hertz (mean-centered) and F values were reported for the MANOVAs. Negative values indicate measures where a reduction was observed in the disgust utterances.

To assist the interpretation of Table 1, one will find that for vowel [ɐ], all three acoustic measures and two of the visual measures differed significantly between disgust and neutral. When compared to neutral, [ɐ] produced in disgust was associated with an average of 13.25Hz CIs [11.00, 15.50] increase in F0, $F(1,807) = 20.48$, $p < .001$, a 25.91Hz CIs [21.5, 30.32] decrease in F1, $F(1,807)=21.44$, $p < .001$ and a 48.83Hz CIs [32.38, 65.28] decrease in F2 $F(1,807)=12.95$, $p < .001$. As for the visual measures, the mixed effects logistic regression found as lip area increases by 1 standard deviation, the likelihood that an utterance was produced in disgust increased by 1.83 ($p < .001$), while a 1 standard deviation decrease in lip width increases the likelihood by a factor of 0.74 ($p < .01$). Table 1 also showed that consistent across all vowels, disgust utterances were produced with significantly higher F0 and lower F1 when compared to neutral expressions.

Table 1. Acoustic and Visual Analysis of Disgust vs Neutral Utterances

Vowel	Measure	Hz dif. or β(Log Odds)	Confidence Intervals	Z or F	Sig.
[ɐ]	F0	13.25	11.00, 15.50	20.48	***
	F1	-25.91	-30.32, -21.25	21.44	***
	F2	-48.83	-65.28, -32.38	12.95	***
	Area	-0.60 (1.83)	-0.85, -0.37	-4.95	***
	Width	0.26 (0.74)	0.08, 0.44	2.64	*
	Height	-	-	-	-
[ɛː]	F0	17.56	11.55, 23.95	15.81	***
	F1	-38.29	-48.07, -28.51	20.90	***
	F2	-	-	-	-
	Area	-	-	-	-
	Width	-	-	-	-
	Height	-0.57 (1.77)	-1.00, -0.18	-4.81	***
[iː]	F0	19.32	15.83, 21.81	16.20	***
	F1	-6.46	-10.90, -2.01	4.53	***
	F2	female −105.86	-153.63, - 57.87	16.74	***
	F2	male −33.89	-81.91, 14.13	-	-
	Area	-	-	-	-
	Width	0.26 (1.41)	0.09, 0.44	3.13	**
	Height	- 0.37 (0.77)	-0.55, -0.19	-4.31	***
[ɔː]	F0	female 28.81	25.70, 39.93	9.44	***
	F0	male 16.54	9.95, 23.89	9.44	***
	F1	female −58.71	-73.39, -44.04	4.53	**
	F1	male −41.72	-56.09, -27.34	4.53	**
	F2	female −105.86	-153.63, -57.87	16.74	***
	F2	male −	-	-	-
	Area	-0.42 (1.35)	-0.64, -0.21	-3.81	***
	Width	0.27 (1.53)	0.10, 0.44	3.13	**
	Height	-0.30 (0.77)	-0.51, -0.09	-2.85	**
[ʊː]	F0	15.02	10.22, 20.24	20.52	***
	F1	−29.85	-38.70, -21.00	2.75	**
	F2	-	-	-	
	Area	-0.46 (1.58)	-0.81, -0.12	-2.63	**
	Width	0.61 (0.54)	0.34, 0.90	4.25	***
	Height	-	-	-	-

Only three vowels ([ɐ] [iː] and [ɔː]) showed a reduction in F2. To illustrate these differences the mean vowel spaces for both neutral and disgust for all speakers is shown in Figure 2. The

vowel space for disgust is shifted such that there is a reduction in F1 and a compression towards the upper boundary of F2.

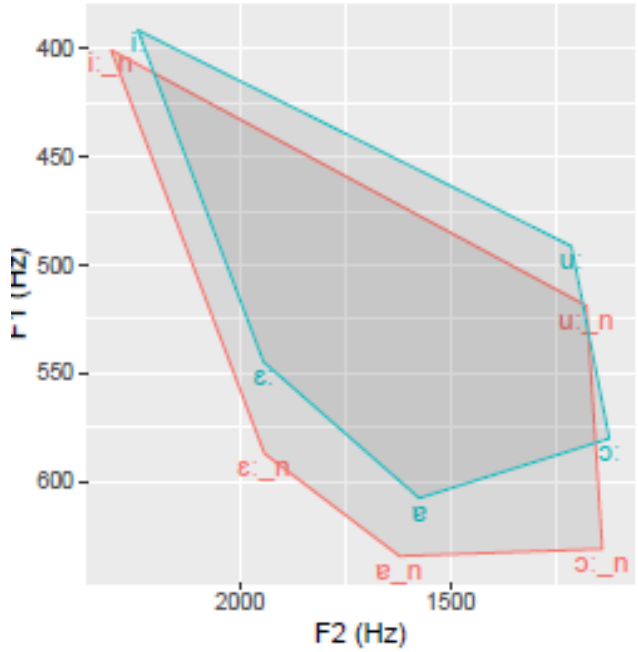

Figure 2: Comparison of the vowel space in a disgust versus neutral tone of voice. The red line and labels appended with '_n' marks the vowel space for of neutral utterances.

It should be noted that where gender differences were observed, Table 1 included separate rows for males and females. In all cases both males and females showed effects in the same direction, with females showing a larger effect than males. The only exception was for the measure of F2 for [ɔ:] which was significant only for the female speakers. There were no interaction effects of lexical tones and emotion types on the examined measures. No gender differences were observed for the visual analysis. While lip area, height and width were generally indicative of disgust, an increase in lip width appeared to be the most salient predictor of disgust.

Discussion

The idea that there are emotion-specific vocal patterns is an old one [23]. However, whether such patterns exist for the expression of disgust has been controversial (e.g., see the view of [24] contra that of [25]). In addressing this issue, our view is that what has been missing in previous research has been a clear proposal concerning how speech gestures and acoustics may be shaped by emotion expressions. More specifically, our interest was to examine how the gestures associated with the expression of disgust affect the acoustic properties of disgust expressive speech. This interest was based on the idea that the gestures that express disgust act as a pathogen avoidance mechanism and we predicted that speech gestures produced in expressing this emotion will have a smaller mouth opening than neutral speech and that this change in mouth configuration may lead to a lowering of the formant frequencies [10]. We compared how the configuration of the mouth (size, height and width) and the acoustics of speech (F0, F1 and F2) changed between disgust and neutral speech in the production of [ɐ], [ɛː], [iː], [ɔː], and [uː] vowels and indeed found evidence that supports our hypothesis.

The observed changes in the acoustic measures showed the impact of the expression of disgust and also confirmed and indicated the articulatory gestures involved. That is, the reduction in F1 across all of the examined vowels likely reflected the reduction in vertical mouth opening and an increase in horizontal mouth opening, such that the retraction of the lips during articulation was more [iː] like, confirming the link of the articulatory and acoustic measures [10]. The observed changes in F2 indicates that disgust expressive speech may also affect tongue configuration such that vowels may be produced with a place of articulation that is further back in the mouth cavity. In line with the idea that the expression of disgust is a pathogen avoidance mechanism, it can be suggested that the tongue may be retracted further back into the mouth cavity to prevent entry into the throat, and also to facilitate expulsion. We are unaware of any studies that have examined the configuration of the tongue position

during disgust expressive speech, so the above suggestion is yet to be tested. However, it is worth pointing out that a transcranial magnetic stimulation study has reported that pictures evoking gustatory disgust suppress the excitability of the tongue representation in the primary motor cortex, suggesting an anticipatory inhibition mechanism preventing the ingestion of contaminants [26].

The robust increase in F0 for the disgust utterances is perhaps surprising given the mixed findings of previous studies. For example, it has been claimed that simulated utterances of disgust are associated with a decrease in F0; whereas those induced by stimuli (such as films) tend to show an increase in F0 [27]. The current study however did not evoke disgust expressions using external stimuli (which would fit the pattern of producing an increase in F0). Thus, it may be that an increase in F0 was due to nonspecific physiological arousal (see ([25]. If the rise in F0 is due to arousal, then in future studies other measures that are sensitive to an individual's physiological status could confirm this. It is worth noting that F0 is not a particularly specific index of emotion due to the range of potential factors that might influence it; this underscores the need for other more pertinent acoustic characteristics of vocal emotion expression (e.g., formant values) and, more importantly, a theory for why any measure should be linked to emotional expression.

This study was able to show that the expression of disgust has an effect on the lip configuration and speech acoustics at the vowel level. In line with previous findings that reported [ɔ:] to be the vowel that is most expressive of disgust; the largest effect sizes observed were associated with this vowel. Further studies should further explore why and how [ɔ:] lends itself to the expression of disgust. Moreover there appears to be vowel specific effects on how formant frequencies are modulated during the expression of disgust. For example, although the reduction in F1 for [i] was significant, the size of the effect was rather small compared to the other vowel types, while F2 was not a significant characteristic of

disgust for [u:] which is naturally produced with low F2. There however appears to be some compensatory effects where vowels that are produced with lower F1 show a larger change in F2 while vowels that are produced with lower F2 showed a reduction in F1 instead. Through such a mechanism, perhaps a greater change in acoustic or vowel quality can more effectively and efficiently convey our expressions.

The only gender differences observed in this study were that the females tended to show larger changes (which were in the same direction) in the acoustic measures than males. This finding appears to coincide with the body of literature that claims that women are more expressive than men [28, 29, 30]. While this study is unable to directly address this question, one explanation is that the expressiveness of speech, or the ease of recognizing an expression increases with the size of changes in acoustic modulation and the exaggeration of articulatory gestures. Further studies may be conducted to examine if the effect size of such changes may disentangle the effects of gender from emotion expressiveness.

Conclusion

The current study focused on a specific emotion (disgust) and indexed articulation by a static measure of mouth height, width and area at the mid-point of vowel articulation. The results supported the specific hypotheses concerning changes in formant values when expressing disgust. What is needed now is a demonstration of how the dynamics of speech production are more generally affected by any ongoing negotiation between properties of speech and other signals that are displayed on the face. Here, the use of continuous measures of jaw, mouth and tongue are needed (e.g., Electromagnetic articulography, EMA) and a broader research agenda undertaken that not only examines the spatial extent of gestures but also their timing.

References

[1] J. M. Tybur, D. Lieberman, and V. Griskevicius, “Microbes, mating, and morality: individual differences in three functional domains of dis- gust.” Journal of personality and social psychology, vol. 97, no. 1, p. 103, 2009.

[2] P. Rozin, J. Haidt, and C. McCauley, “Disgust,” Hand- book of Emotions, pp. 757–776, 2008.

[3] V. Curtis and A. Biran, “Dirt, disgust, and disease: Is hygiene in our genes?” Perspectives in biology and medicine, vol. 44, no. 1, pp. 17–31, 2001.

[4] P. Ekman, W. Friesen, M. O’ Sullivan, I. Diacoyanni-Tarlatzis, R. Krause, T. Pitcairn, K. Scherer, A. Chan, K. Heider, W. LeCompte, P. RicciBitti, and M. Tomita, “Universals And Cultural Differences In The Judgment Of Facial Expressions of Emotion,” Journal of Personality and Social Psychology, vol. 53, no. 4, pp. 712–717, 1987.

[5] J. M. Susskind, D. H. Lee, A. Cusi, R. Feiman, W. Grabski, and A. K. Anderson, “Expressing fear enhances sensory acquisition.” Nature neu- roscience, vol. 11, no. 7, pp. 843–850, 2008.

[6] D. Fessler and K. Haley, “Guarding the perimeter: The outside- inside dichotomy in disgust and bodily experience,” Cognition & Emotion, vol. 20, no. 1, pp. 3–19, 2006. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02699930500215181

[7] P. Rozin, L. Lowery, and R. Ebert, “Varieties of disgust faces and the structure of disgust,” Journal of personality and social psychology, vol. 66, no. 5, pp. 870–881, 1994.

[8] G. Bailly, A. B´egault, F. Elisei, and P. Badin, “Speaking with smile or disgust : data and models,” AVSP, pp. 111–116, 2008.

[9] E. M. Caldognetto, P. Cosi, C. Drioli, G. Tisato, and F. Cavicchio, “Modifications of phonetic labial targets in emotive speech: effects of the co-production of speech and emotions,” Speech Communication, vol. 44, no. 1, pp. 173–185, 2004.

[10] E. F. Lindblom and E. F. Sundberg, “Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement,” The Journal of the Acoustical Society of America, vol. 50, no. 4, 1971.

[11] V. C. Tartter, “Happy talk: Perceptual and acoustic effects of smiling on speech.” Perception & psychophysics, vol. 27, no. 1, pp. 24–27, 1980.

[12] V. C. Tartter and D. Braun, “Hearing smiles and frowns in normal and whisper registers,” The Journal of the Acoustical Society of America, vol. 96, no. 4, pp. 2101–2107, 1994.

[13] P. N. Juslin and P. Laukka, “Communication of emotions in vocal ex- pression and music performance: Different channels, same code?” Psy- chological bulletin, vol. 129, no. 5, p. 770, 2003.

[14] L. Kaiser, “Communication of Affects by Single Vowels,” Synthese, vol. 14, no. 4, pp. 300–319, 1962.

[15] C. S. Chong, J. Kim, and C. Davis, “The sound of disgust: How facial expression may influence speech production,” Interspeech, pp. 37–41, 2016.

[16] R. E. Shor, “The production and judgment of smile magnitude,” The Journal of General Psychology, vol. 98, no. 1, pp. 79–96, 1978.

[17] C. S. Chong, J. Kim, and C. Davis, “Development of an Audiovisual Cantonese Emotional Speech Database,” The 17th conference of the Oriental chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment /CASLRE (Confer- ence on Asian Spoken Language Research and Evaluation), 2014.

[18] L. L. Wong and S. D. Soli, “Development of the cantonese hearing in noise test (chint),” Ear and hearing, vol. 26, no. 3, pp. 276–289, 2005.

[19] E. Velten, “A laboratory task for induction of mood states,” Behaviour research and therapy, vol. 6, no. 4, pp. 473–482, 1968.

[20] J. Wilting, E. Krahmer, and M. Swerts, “Real vs acted emotional speech,” in Interspeech, 2006, pp. 805–808.

[21] Y. Xu, “Prosodyproa tool for large-scale systematic prosody analysis.” Laboratoire Parole et Langage, France, 2013.

[22] D. Bates, M. Maechler, B. Bolker, S. Walker et al., “lme4: Linear mixed- effects models using eigen and s4,” R package version, vol. 1, no. 7, pp. 1–23, 2014.

[23] C. Darwin, The expression of the emotions in man and animals, 1872. [24] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, “Factors in the recognition of vocally expressed emotions: A comparison of four languages,” Journal of Phonetics, vol. 37, no. 4, pp. 417–435, 2009.

[24] Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, vol. 37, Issue 4, pp. 417-435

[25] R. Banse and K. R. Scherer, “Acoustic profiles in vocal emo- tion expression.” Journal of personality and social psychology, vol. 70, no. 3, pp. 614–36, 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8851745

[26] C. M. Vicario, R. D. Rafal, S. Borgomaneri, R. Paracampo, A. Kri- tikos, and A. Avenanti, “Pictures of disgusting foods and disgusted fa- cial expressions suppress the tongue motor cortex,” Social cognitive and affective neuroscience, vol. 12, no. 2, pp. 352–362, 2016.

[27] K. R. Scherer, “Non-linguistic indicators of emotion and psychopathol- ogy,” in Emotions in personality and psychopathology, C. E. Izard, Ed., New York: Plenum Press, pp. 495–529, 1979.

[28] A. M. Kring and A. H. Gordon, “Sex differences in emotion: expression, experience, and physiology.” Journal of personality and social psychol- ogy, vol. 74, no. 3, p. 686, 1998.

[29] R. D. Ashmore and F. K. Del Boca, “Sex stereotypes and implicit per- sonality theory: Toward a cognitivesocial psychological conceptualiza- tion,” Sex roles, vol. 5, no. 2, pp. 219–248, 1979.

[30] L. R. Brody and J. A. Hall, “Gender, emotion, and expression,” Hand- book of emotions, vol. 2, pp. 338–349, 2000.

Chapter 5. Lip Measurement Algorithm

The accurate detection and segmentation of lips in images has attracted a lot of attention in the computer vision, image processing (Xing, Sieber, & Kalacska, 2014) and automated speech recognition communities (Chan, 2001; Erber, 1969; Kaynak, Zhi, Cheok, Sengupta, & Chung, 2001; Rabi & Lu, 1998; Zhang, Levinson & Huang, 2000). This interest stems from the wide range of applications where visual information from the lips can improve the performance and robustness of systems of automated audio-visual speech and facial expression recognition (Erber, 1969; Kaynak, Zhi, Cheok, Sengupta, & Chung, 2001; Rabi & Lu, 1998; Zhang, Levinson & Huang, 2000; Happy & Routray, 2015).

Although many different methods and systems for lip segmentation have been proposed and developed (Eveno, Caplier, & Coulon, 2001; Leung, Wang, & Lau, 2004; Liévin, & Luthon, 1999; Liew, Leung & Lau, 2003; Wang, Lau, Liew, & Leung, 2007), these methods are typically finely tuned to the training database and are not robust to variations in individual lip configurations, different lip and skin colour tones, lighting conditions, presence of facial hair and so forth which can significantly affect performance. Moreover, many of these methods and systems impose certain constraints on the users such as wearing a head mounted camera or painting the subject's lips (see Meier, Stiefelhagen, Yang, & Waibel, 2000) thus precluding application on our stimuli.

In this chapter I present my research efforts towards developing a lip segmentation algorithm that was used in the experiment presented in chapter 5. This chapter is structured as a methods chapter and is intended to provide the reader with a broad understanding of the general rationale and aim of each step of the algorithm.

The developed algorithm combines edge detection and colour segmentation, which are techniques that are commonly used for lip segmentation tasks. Some of the earliest work on lip segmentation used edge detection (Zhang & Mersereau, 2000), which is an image processing technique for detecting discontinuities in brightness which indicates the boundaries of objects. Colour information is also widely used in lip detection tasks (Duchnowski, Hunke, Busching, Meier, & Waibel, 1995; Hsu, Abdel-Mottaleb, & Jain, 2002; Sadeghi, Kittler, & Messer, 2002) where segmentation can be achieved through the differences in the colour features of image pixels (e.g., lips and skin have different concentrations of green). Other techniques include building models of the lips such as, snakes (Delmas, Coulon, & Fristot, 1999), active contour models (Liu, Cheung, Li, & Liu, 2010) and deformable templates (Liew, Leung, & Lau, 2002).

Techniques that require the building of a lip model were not used in this algorithm as the construction of these models is often very challenging and time consuming; and are thus unsuitable for the scope and purposes of this thesis. It should also be noted that the algorithm was designed specifically for the purposes of this thesis and is not intended to be an improvement over pre-existing techniques. The algorithm is implemented in Matlab (2016) using functions from the Image Processing Toolbox (version 9.5) and is appended as Appendix 1. The following section provides a description of each of the steps in the algorithm.

The algorithm

Image frames of interest were imported into Matlab and the heights of the images were cropped to retain only the area below the speakers' philtrum. The removal of the nostrils and other irrelevant facial features (e.g., eyes and nose) reduces the complexity of the image and the amount of data to be processed (see Figure 1).

Figure 1. A sample cropped image frame.

A texture map of the image was created through the application of an entropy filter. The texture map reduces the complexity of the image by examining the variation in intensities of surrounding each pixel of the image. The texture map is therefore a matrix of indices that contains the entropy value or the variability of intensity values of the 9-by-9 pixel neighbourhood around each pixel. A pixel with low entropy values suggests that there is low variability in intensity around the pixel and it therefore likely has the same textural characteristic as its neighbouring pixels. In contrast, a pixel with high entropy value suggests that there is large variation in intensities around and pixel, hence the pixel is likely to be an "edge" that borders regions with different textures or objects. Figure 2 shows an example of a texture map. Note that the texture map does not distinguish the lip region well and has introduced noise to the image by highlighting creases in the background.

Figure 2. A sample output after applying the entropy filter.

In order to sharpen the detected boundaries, the contrast of the image was enhanced using the histogram equalisation method. Histogram equalisation is a mathematical method that rescales the image intensity by expanding and redistributing the pixel intensity range of an image. The figures in 3 show the increase in contrast after histogram equalisation. Note that pixels with the largest entropy values (strongest edge boundaries) are the brightest pixels.

A thresholding function was then applied to remove boundaries with low contrast and to convert the image to a binary image (only white or black pixels). Pixels with intensity values less than one standard deviation above the mean were removed (converted to black). All remaining pixels were converted to white. Regions that were enclosed by edge boundaries were filled in resulting in the image shown in Figure 4. This reduced the amount of noise in the image, retaining only the regions with the strongest edge boundaries.

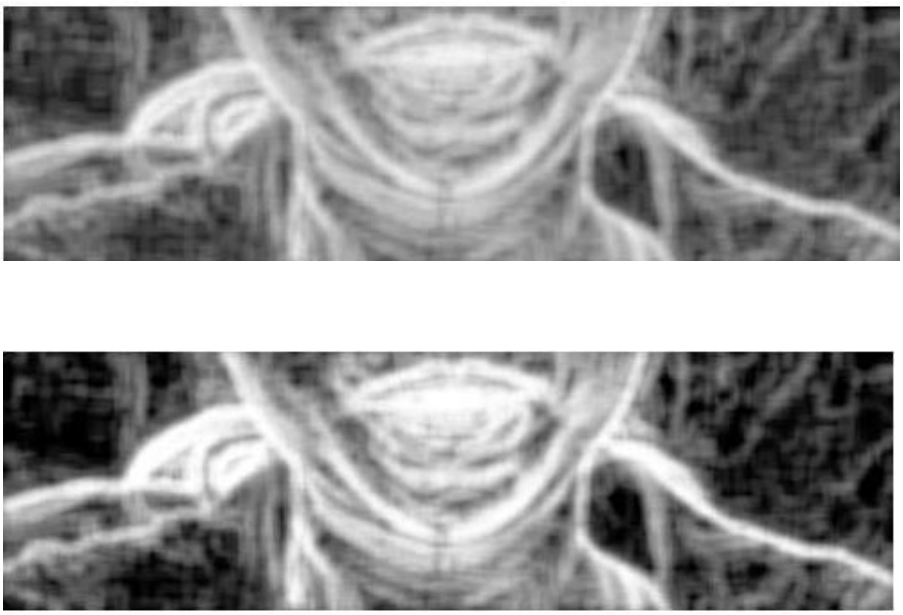

Figure 3. Before (top) and after (bottom) histogram equalisation.

As the image still included irrelevant regions (e.g., hair and jawline), the image was further refined by applying a second thresholding function that removed all regions with an area of less than 900 pixels (this value was determined to be suitable for the stimuli used). This resulted in Figure 5.

Figure 4. The image above shows regions in the image with the highest contrast.

Figure 5. Figure 4 after removing all regions that were less than 900 pixels in size.

While the second thresholding function is generally adequate in removing all but the lip region, as with the example shown in Figure 5, there are scenarios where more than one region remains. In these cases, the region whose centroid is closest to the middle of the upper boundary of the image is retained. This region most likely corresponds to the lips due to the image cropping procedure.

Figure 6 below shows the solution of the edge segmentation method. While this method performed well in detecting the cupid bow feature of the upper lips, it was unable to accurately detect the vermillion border of the lower lip. This is a common shortcoming of the

edge detection method which underperforms when there is a lack of a strong boundary in the lower lip region due to 1) the low textural contrast between the vermillion border and surrounding skin and 2) poorer illumination.

Figure 6. The solution of the edge detection technique.

To compensate for the above shortcoming, a colour segmentation method was applied. This method is based on the principle that the intensity of the colour green is less in the pigmentation of the lips than in the skin. Using the solution obtained from the texture segmentation method, a bounding box was applied to crop the lip region of the image. The cropped image was then converted into the HSV (hue, saturation and value) colour space and the saturation (the intensity or shade of colour) of the image was maximised. An index of the difference between the red and green components for each pixel of the image was calculated using the formula below.

$$\text{Index} = (\text{red} - \text{green}) + (\text{blue} - \text{green})$$

Pixels with an index value greater than one standard deviation above the mean were retained. Figure 7 below shows the solution.

Figure 7. Lip segmentation using colour transformation

Contrary to the edge detection method, colour transformation performed better at detecting the vermillion boundary of the lower lip than the upper lip. Hence, the final step combined the solution from both methods; the solution from the edge detection technique for the upper lip was merged with the solution from the colour transformation method for the lower lip. The final output can be seen in Figure 8 below.

Figure 8. Final lip segmentation solution.

For the purposes of the experiment presented in the previous chapter, the measure of area was defined as the number of pixels contained within the segmented region. The measures of lip height and length were derived as the maximum height and length of the segmented region. For validation purposes, 50 images were randomly selected for manual measurement by hand. Paired samples Wilcoxon tests indicated that the manual measures did not differ significantly from the algorithm outputs.

While the algorithm performed well for the majority of our stimuli, lip segmentation is a difficult task and there are several limitations of our algorithm that should be noted. One, this algorithm was developed specifically for the purposes of this thesis and the CAVES database and is not intended to be a robust technique that is applicable to other databases. Two, the validity of this algorithm has not been evaluated outside of the context of this thesis, nevertheless, the algorithm adopted different techniques that are widely used and validated in other applications. Three, the output of this algorithm is most suited for within group analyses such as the ones conducted in the previous chapter due to the individualised tuning of parameters.

References

Chan, M. T. (2001). HMM-based audio-visual speech recognition integrating geometric-and appearance-based visual features. In *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on* (pp. 9-14). IEEE.

Delmas, P., Coulon, P. Y., & Fristot, V. (1999). Automatic snakes for robust lip boundaries extraction. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on* (Vol. 6, pp. 3069-3072). IEEE.

Duchnowski, P., Hunke, M., Busching, D., Meier, U., & Waibel, A. (1995). Toward movement-invariant automatic lip-reading and speech recognition. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on* (Vol. 1, pp. 109-112). IEEE.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech, Language, and Hearing Research*, *12*(2), 423-425.

Eveno, N., Caplier, A., & Coulon, P. Y. (2001). New color transformation for lips segmentation. In *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on* (pp. 3-8). IEEE.

Happy, S. L., & Routray, A. (2015). Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, *6*(1), 1-12.

Hsu, R. L., Abdel-Mottaleb, M., & Jain, A. K. (2002). Face detection in color images. *IEEE transactions on pattern analysis and machine intelligence*, *24*(5), 696-706.

Kaynak, M. N., Zhi, Q., Cheok, A. D., Sengupta, K., & Chung, K. C. (2001). Audio-visual modeling for bimodal speech recognition. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on* (Vol. 1, pp. 181-186). IEEE.

Leung, S. H., Wang, S. L., & Lau, W. H. (2004). Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE transactions on image processing*, *13*(1), 51-62.

Liévin, M., & Luthon, F. (1999). Unsupervised lip segmentation under natural conditions. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'99)* (pp. 3065-3068).

Liew, A. W. C., Leung, S. H., & Lau, W. H. (2002). Lip contour extraction from color images using a deformable model. *Pattern Recognition*, *35*(12), 2949-2962.

Liew, A. C., Leung, S. H., & Lau, W. H. (2003). Segmentation of color lip images by spatial fuzzy clustering. *IEEE transactions on Fuzzy Systems*, *11*(4), 542-549.

Liu, X., Cheung, Y. M., Li, M., & Liu, H. (2010). A lip contour extraction method using localized active contour model with automatic parameter selection. In *Pattern Recognition (ICPR), 2010 20th International Conference on* (pp. 4332-4335). IEEE.

Meier, U., Stiefelhagen, R., Yang, J., & Waibel, A. (2000). Towards unrestricted lip reading. *International Journal of Pattern Recognition and Artificial Intelligence*, *14*(05), 571-585.

Rabi, G., & Lu, S. W. (1998). Visual speech recognition by recurrent neural networks. *Journal of Electronic Imaging*, *7*(1), 61-70.

Sadeghi, M., Kittler, J., & Messer, K. (2002). Modelling and segmentation of lip area in face images. *IEE Proceedings-Vision, Image and Signal Processing*, *149*(3), 179-184.

Wang, S. L., Lau, W. H., Liew, A. W. C., & Leung, S. H. (2007). Robust lip region segmentation for lip images with complex background. *Pattern Recognition*, *40*(12), 3481-3491.

Xing, J., Sieber, R., & Kalacska, M. (2014). The challenges of image segmentation in big remotely sensed imagery data. *Annals of GIS*, *20*(4), 233-244.

Zhang, Y., Levinson, S., & Huang, T. (2000). Speaker independent audio-visual speech recognition. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on* (Vol. 2, pp. 1073-1076). IEEE.

Zhang, X., & Mersereau, R. M. (2000). Lip feature extraction towards an automatic speechreading system. In *Image Processing, 2000. Proceedings. 2000 International Conference on* (Vol. 3, pp. 226-229). IEEE.

General Discussion

The thesis examined how characteristics of spoken language and facial expressions affect the production of expressive speech by investigating how spoken expression of emotions are produced and perceived by Cantonese (tone language) and English (non-tone language) speakers (Chapter 1-3), and how facial expression of disgust affects speech articulation and the F1 and F2 properties of speech (chapter 4-5). The important findings and implications in relation to language and facial expressions are discussed separately in the following sections.

Language specific differences

To investigate the influence of spoken language on spoken emotions perception, I conducted a perception study in which three groups of participants (Australian English, Malaysian Malay and Hong Kong Cantonese speakers) identified spoken expression of emotions that were produced in either English or Cantonese (Chapter 2). The results showed that those who spoke a similar type of language (e.g., non-tone language, Malay and English), as opposed to having a similar culture (Malay and Cantonese), were more similar in how they perceived expressive speech. This result demonstrates that the characteristics of spoken language can have an influence on the perception of expressive speech.

Following up on the above findings, I conducted a production study by analysing the acoustic characteristics of expressive speech in English and Cantonese. I found that compared to English, Cantonese expressive speech was produced with lower median F0 and flatter F0 contour. These results are aligned with an earlier study which reported that spoken expression of emotions produced in Mandarin used a smaller F0 frequency range when compared to Italian (Anolli, Wang, Mantovani & De Toni, 2008). These results support the claim that tone language users may restrict the use of F0 cues in the production of spoken expressions of emotions in order to preserve F0 as a linguistic resource to convey semantic meaning (Ross,

Edmondson & Seibert, 1986; Anawin, 1998; Wang & Lee, 2015). Furthermore, the results of my study showed that, when compared to English, Cantonese expressive speech not only used fewer F0 cues, but used these cues to code emotions in a systematically different way. That is, while the same types of F0 cues helped distinguish spoken expressions, these cues (e.g. low median F0 and high number of turning points) appeared to code different emotions depending on whether the expression was produced in Cantonese or English.

Taken together, the findings of the above production and perception studies demonstrate that there are systematic differences in how tone and non-tone languages speakers produce and perceive expressive speech. This suggests that if we view emotion as a form of information that is transmitted by the speech signal, the manner in which this information is conveyed may differ between groups depending on the characteristics of the spoken language.

Of course, the current investigation is limited to only a single tone (Cantonese) and non-tone language (English); and to Cantonese, English and Malay participant groups. The findings will therefore be strengthened if additional perception and acoustic analysis studies included other tone and non-tone languages and additional complementary groups of participants. For example, the inclusion of another group of tone language speakers (Mandarin for example) can be useful in further demonstrating if there were greater perceptual similarities among those who spoke a similar type of language. A more balanced research design may also be achieved if the perception study included a tone language speaker group that has greater cultural similarity to the Australian speakers (individualistic culture) as a contrast to the Malay speakers.

Furthermore, the current investigation examined only a single acoustic property, i.e., F0, thus it is unclear whether other acoustic cues may be affected by the differences in how F0 is used in the production of spoken expression of emotions in tone languages. In order to address

these limitations, I intend to conduct further acoustic analysis and machine learning studies, extending the examination to include a greater range of acoustic features. In particular, I aim to examine if a greater weight may be placed on other acoustic cues such as speech rate and intensity possibly as a way of compensating for the restrictions on F0 use in the production of emotions in tone languages. I also aim to examine a larger selection of languages to determine if the results of this thesis may be generalisable to other tone and non-tone languages. In conducting this work, I aim develop expressive speech databases in other language such as Malay using a similar design of the CAVES (i.e., a large selection of sentences that were repeated in different emotion types).

Facial expressions

In Chapter 4, I demonstrated that compared to neutral speech, spoken expressions of disgust were produced with smaller vertical mouth opening, greater horizontal mouth opening and lower F1 and F2 frequencies. When viewed against the findings that spoken expression of emotions may differ between groups due to language specific differences, the results of Chapter 4 hint at aspects of expressive speech that may be language universal. That is, while the use of F0 may vary as a function of the characteristics of spoken language, acoustic features such as F1 and F2 may be used in a similar manner across languages when producing disgust due to the adaptive functional role of its facial expression as a pathogen avoidance mechanism.

The current thesis experiments were however limited to only a single language (Cantonese) and a single facial expression (disgust). The next step would be to examine if changes to articulation and the acoustic properties of F1 and F2 in expressions of disgust may be observed in other languages. In addition, the measurements of lip configuration may be conducted using more sophisticated methods such as 3D motion capture techniques using

systems such as Optotrack or Vicon. I chose to use the lip segmentation algorithm in this thesis despite of its limitations as 3D motion capture methods generally require the application of sensors or markers to the face which obscures the face thereby limiting the utility of the database for audio-visual emotion research and the naturalness of the expressions (although non-invasive infra-red based tracking systems, such as those used by Apple show promise).

In conducting further investigations, considerations will be given to the use of 3D capture methods to further illuminate the range of interaction between facial expressions, speech articulation and speech acoustics. I also aim to examine how other facial expressions of emotions such as smiles (see Tartter, 1980) may have an impact on the articulation of expressive speech.

Concluding remarks

The findings of the thesis highlights some of the factors that may contribute to intergroup differences in emotion perception while also offering some insights into how aspects of expressive speech may be shaped by language specific and language universal factors. Further work in search of acoustic correlates of emotions may need to adopt an auditory-visual approach while staying mindful of how the unique characteristics of the examined language may affect expressive speech.

References

Adolphs, R., & Anderson, D. J. (2018). *The neuroscience of emotion: A new synthesis*. Princeton University Press.

Anawin, L. (1998). Intonation in Thai. Intonation Systems: A Survey of Twenty Language, Edited by D. Hirst and AD Cristo, 376-394.

Anolli, L., Wang, L., Mantovani, F., & De Toni, A. (2008). The voice of emotion in Chinese and Italian young adults. Journal of Cross-Cultural Psychology, 39(5), 565-598.

Bailly, G., Bégault, A., Elisei, F., & Badin, P. (2008). Speaking with smile or disgust: data and models. In Auditory-Visual Speech Processing (AVSP) (pp. 111-116).

Caldognetto, E. M., Cosi, P., Drioli, C., Tisato, G., & Cavicchio, F. (2004). Modifications of phonetic labial targets in emotive speech: effects of the co-production of speech and emotions. *Speech Communication*, *44*(1-4), 173-185.

Cannon, W. B. (1916). Bodily changes in pain, hunger, fear, and rage: An account of recent researches into the function of emotional excitement. D. Appleton.

Chapman, H. A., Kim, D. A., Susskind, J. M., & Anderson, A. K. (2009). In bad taste: Evidence for the oral origins of moral disgust. *Science*, *323*(5918), 1222-1226.

Collins, E. F., & Bahar, E. (2000). To know shame: Malu and its uses in Malay societies. *Crossroads: An Interdisciplinary Journal of Southeast Asian Studies*, 35-69.

Curtis, V., Aunger, R., & Rabie, T. (2004). Evidence that disgust evolved to protect from risk of disease. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(suppl_4), S131-S133.

Curtis, V., De Barra, M., & Aunger, R. (2011). Disgust as an adaptive system for disease avoidance behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1563), 389-401.

Damasio, A. (2019). *The Strange Order of Things: Life, Feeling, and the Making of Cultures*. Vintage.

Eibl-Eibesfeldt, I. (2017). *Human ethology*. Routledge.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169-200.

Elfenbein, H. A. (2013). Nonverbal dialects and accents in facial expressions of emotion. *Emotion Review*, *5*(1), 90-96.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, *128*(2), 203.

Elfenbein, H. A., & Ambady, N. (2003). Universals and cultural differences in recognizing emotions. *Current directions in psychological science*, *12*(5), 159-164.

Goddard, C. (1996). The" social emotions" of Malay (Bahasa melayu). ETHOS-BERKELEY-UNIVERSITY OF CALIFORNIA THEN WASHINGTON DC-, 24, 426-464.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code?. *Psychological bulletin*, *129*(5), 770.

Keltner, D., & Gross, J. J. (1999). Functional accounts of emotions. *Cognition & Emotion*, *13*(5), 467-480.

Kim, U., Triandis, H. C., Kagitcibasi, C., Choi, S. C., & Yoon, G. (Eds.). (1994). Individualism and collectivism: Theory, method, and applications. Thousand Oaks, CA: Sage.

LeDoux, J. (2012). Rethinking the Emotional Brain. *Neuron*, *73*.

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

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological review*, 98(2), 224.

Matsumoto, D. (1992). American-Japanese cultural differences in the recognition of universal facial expressions. *Journal of cross-cultural psychology*, *23*(1), 72-84.

Matsumoto, D. (1999). American-Japanese cultural differences in judgements of expression intensity and subjective experience. *Cognition & Emotion*, *13*(2), 201-218.

Matsumoto, D., Yoo, S. H., & Fontaine, J. (2008). Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of cross-cultural psychology*, 39(1), 55-74.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*(2), 107-120.

Ross, E. D., Edmondson, J. A., & Seibert, G. B. (1986). The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice. *Journal of phonetics*.

Rozin, P., & Fallon, A. E. (1987). A perspective on disgust. *Psychological review*, *94*(1), 23.

Rozin, P., Haidt, J., & McCauley, C. R. (2008). *Disgust*.

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language*, *27*(1), 40-58.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural psychology*, *32*(1), 76-92.

Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and emotion*, *1*(4), 331-346.

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Shariff, A. F., & Tracy, J. L. (2011). What are emotion expressions for?. *Current Directions in Psychological Science*, *20*(6), 395-399.

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological science*, *16*(3), 184-189.

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118*(5), 3177-3186.

Susskind, J. M., Lee, D. H., Cusi, A., Feiman, R., Grabski, W., & Anderson, A. K. (2008). Expressing fear enhances sensory acquisition. *Nature neuroscience*, *11*(7), 843.

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & psychophysics*, *27*(1), 24-27.

Thompson, W. F., & Balkwill, L. L. (2006). Decoding speech prosody in five languages. *Semiotica*, *2006*(158), 407-424.

Tybur, J. M., Lieberman, D., & Griskevicius, V. (2009). Microbes, mating, and morality: individual differences in three functional domains of disgust. *Journal of personality and social psychology*, 97(1), 103.

Wang, T., & Lee, Y. C. (2015). Does restriction of pitch variation affect the perception of vocal emotions in Mandarin Chinese?. *The Journal of the Acoustical Society of America*, 137(1), EL117-EL123.

Yamaguchi, S. (1994). Collectivism among the Japanese: A perspective from the self. In U. Kim & H. C. Triandis (Eds.), *Individualism and collectivism: Theory, method, and applications* (Vol. 18, pp. 175-188). Thousand Oaks, CA: Sage.

Appendix 1. Lip Segmentation Algorithm

```
% Image directory/input folder[20]
root = 'C:\Users\Images\';
flist=dir([root '*.png']);

% y-axis location of the middle of the philtrum.
phil=360;

% location of output/output folder
picpath=[root 'lips\'];
        if ~exist (picpath)
            mkdir (picpath)
        end

% Create data structure
Data=struct('Vid', [], 'Emotion',[], 'Vowel', [], 'Area', [], 'Length', [], 'Height',[]);
iD=1;

% Loop through all images in the input folder
for iFile=1:length(flist)
            file=flist(iFile).name;

% Read image
f=imread([root file]);

% Crop to include only regions below the philtrum
```

[20] Comments are colored in blue

```
  m = f(phil:end,:,:);

% Apply Entropy Filter and Histogram Equalisation
E=entropyfilt(m);
Eim=mat2gray(E);
Eim1=rgb2gray(Eim);
tt=histeq(Eim1);

% Apply thresholding on Entropy Map
tt1=tt(:);
mtt=mean(tt1);
stt=std2(tt1);
ttTresh=mtt+sttt+.1;
tt(tt<ttTresh)=0;
tt(tt>.0)=255;

% Create filled regions and remove regions smaller than 900 pixels in size
bwfill=imfill(tt,'holes');
bwfill=bwareaopen(bwfill,900);

% Check if there is more than 1 region remaining and keep only the region with centroid
closest to the top centre of the image
[L,num]=bwlabel(bwfill);

if num > 1
        stats=regionprops(L,'Area' , 'Centroid');
        Obscentroids=cat(1,stats.Centroid);

                for
                iCent = 1:length(Obscentroids)
```

```
            Dist{iCent} = (abs(Obscentroids(iCent) - (size)/2)) + (abs(Obscentroids(iCent,2) - 25));

            end


        iDealCent=find(cell2mat(Dist) == min(cell2mat(Dist)));

        gg=find(L ~= iDealCent);

        BWfilt = bwfill;

        BWfilt(gg)=0;

        Dist=[];

else

        BWfilt=bwfill;

end


%Print image for visual inspection

BWoutline1 = bwperim(BWfilt);

Segout =m;

 Segout(BWoutline1) = 255;

figure, imshow(Segout), title('texture');


% Get measures from entropy filter solution and crop lip region for colour transformation method

stats=regionprops(BWfilt,'BoundingBox');

mouth=m(stats.BoundingBox(2) -5:(stats.BoundingBox(2) + stats.BoundingBox(4) + 20), stats.BoundingBox(1) : (stats.BoundingBox(1) + stats.BoundingBox(3)),: );


% Colour transformation / detection

Im=rgb2hsv(f);

Im(:,:,2)=f(:,:,2)*2;

Im(Im > 1) = 1;

Im=hsv2rgb(Im);
```

```
x1=stats.BoundingBox(2)-20;
x2=(stats.BoundingBox(2)+stats.BoundingBox(4))+50;
y1=stats.BoundingBox(1)-50;
y2=(stats.BoundingBox(1)+stats.BoundingBox(3)+50);

xx=round((x2-x1)/2);
yy=round((y2-y1)/2);

x1=phil+x1;
x2=phil+x2;
 Im=Im(x1:x2,y1:y2,: );

        r=Im(:,:,1);
        g=Im(:,:,2);
        b=Im(:,:,3);

% Thresholding
Im1=(r-g) + (b-g);
ii=Im1(:);
ave=mean(ii);
sd=std2(ii);
threshold=ave+sd-.02;

gray=rgb2gray(Im);
imTr=gray;
imTr(I<threshold)=0;
imTr(imTr>0)=255;
imTr = imfill(imTr, 'holes');
```

```
[L,num]=bwlabel(imTr);

if num > 1

    imTr=bwareaopen(imTr,800);
    [L,num]=bwlabel(imTr);

    if num > 1
      stats=regionprops(L,'Area' , 'Centroid');
      Obscentroids=cat(1,stats.Centroid);

          for iCent = 1:length(Obscentroids)
              Dist{iCent} =  (abs(Obscentroids(iCent) - yy)) + (abs(Obscentroids(iCent,2) - xx));
          end

      iDealCent=find(cell2mat(Dist) == min(cell2mat(Dist)));
      gg=find(L ~= iDealCent);
      BWfilt = imTr;
      BWfilt(gg)=0;
      Dist=[];
    else
       BWfilt=imTr;
    end

end

lowerlip=BWfilt(16:end-30,51:end-50);
[L,num]=bwlabel(lowerlip);
```

```
if num > 1
    stats1=regionprops(L,'Area');
    roi=max([stats1.Area]);
    filt=roi-1;
    lowerlip=bwareaopen(lowerlip,filt);
end

% Find extrema and merge entropy and colour transform solutions
 s=regionprops(L,'Extrema');
  leftop=s.Extrema(8,2);
 rightop=s.Extrema(3,2);
 ss=round(max(leftop,rightop));

 upperlip=BWfinal(1:ss-1,:);
 llowerlip=lowerlip(ss:end,:);

 BinMask=vertcat(upperlip,llowerlip);
  se90=strel('line',2,90);
se0=strel('line',2,0);
BinMask1 = imdilate(BinMask, [se90 se0]);
BinMask1=bwareaopen(BinMask1,50);
 BinMask2 = imfill(BinMask1, 'holes');

BWoutline1 = bwperim(BinMask2);
Segout =mouth;
Segout(BWoutline1) = 255;

 %Save output
 imwrite(Segout, [picpath flist(iFile).name]);
```

```
%Extract measures and save in data structure
 II=regionprops(BWfinal,'Area','BoundingBox');

                    Data.Vid{iD}=file(1:end-4);
                    strs=strsplit(file, '_');
                    Data.Vowel{iD}=strs{4};
                    Data.Emotion{iD}=strs{2};
                    Data.Length(iD)=II.BoundingBox(1) + II.BoundingBox(3);
                    Data.Height(iD)=II.BoundingBox(2) + II.BoundingBox(4);
                    Data.Area(iD)=II.Area;
                    iD=iD+1;

          end

%Write data as .txt file
fout=[root 'lips.txt'];
Dwrite(Data, fout, '%s %s %s %f %f %f ')
```