Speech Communication xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Disgust expressive speech: The acoustic consequences of the facial expression of emotion

Chee Seng Chong*, Jeesun Kim, Chris Davis

The MARCS Institute, Western Sydney University, 2 Bullecourt Avenue, Milperra, Sydney, 2214, Australia

ARTICLE INFO

Keywords:
Disgust
Acoustic profile of emotions
Expressive speech
Formants

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], [ɛː], [iː], [ɔː], [uː]) 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.

1. 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 (Tybur et al., 2009; Rozin et al., 2008) and thus the types of stimuli that elicit disgust (Curtis and Biran, 2001) and the way it is expressed is similar across cultures (Ekman et al., 1987). 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 (Susskind et al., 2008; Fessler and Haley, 2006; Rozin et al., 1994).

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 stereotyped 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 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 (Bailly et al., 2008). Looking more specifically at lip configurations using reflective markers, a study by Caldognetto et al. (2004) 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

* Corresponding author.
E-mail address: l.chong@westernsydney.edu.au (C.S. Chong).

https://doi.org/10.1016/j.specom.2017.12.007
Received 22 January 2017; Received in revised form 28 November 2017; Accepted 15 December 2017

of the first formant (*F1*) (Lindblom and Sundberg, 1971), 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 (Tartter, 1980). 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 (Tartter and Braun, 1994). 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 (Juslin and Laukka, 2003), only one study has examined this possibility (Kaiser, 1962). 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 (*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 previously (Chong et al., 2016). In a preliminary study where we compared disgust and neutral speech of five female speakers, we found the vowel [ɔː] to have the largest changes in formant frequencies. We suggested that open and rounded vowels such as [o] 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 (Chong et al., 2016). 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 (Tartter, 1980), the assumption between smiling and increased mouth opening during speech production was made based on a study of non-verbal smiles by Shor (1978). Likewise, the relationship between lip configuration and the expression of disgust was an untested assumption in Chong et al. (2016). 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 (Chong et al., 2016), 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], [ɛː], [iː], [ɔː], [uː]) 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.

2. Methods

2.1. Material

Speech materials were obtained from the Cantonese Audio-Visual Expressive (CAVE) speech database (Chong et al., 2014) 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 (Wong and Soli, 2005) 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 × 50 sentences × 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. Emotion induction was carried out using a modification to the Velten method (Velten, 1968) similar to Wilting et al. (2006). In this case, the speakers first read a scenario to induce the targeted emotional state. They then 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.

2.2. Acoustic measure

This analysis is similar to what was reported in Chong et al. (2016). The Cantonese sentences were first transcribed into Jyutping (a Romanisation system for Cantonese,[1]) 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 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 (Xu, 2013), a script implemented in Praat. Mean formant values were extracted using Linear Predictive Coding (LPC) analysis implemented through Burg's algorithm in Praat.

2.3. 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. Fig. 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.

Due to the amount of data and to minimize human error, the above steps were automated using a Matlab script. 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 Fig. 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

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