INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The sound of disgust: how facial expression may influence speech production

Chee Seng Chong, Jeesun Kim, Chris Davis

The MARCS Institute, Western Sydney University, Australia

L.chong@westernsydney.edu.au, j.kim@westernsydney.edu.au, chris.davis@westernsydney.edu.au

Abstract

In speech articulation, mouth/lip shapes determine properties of the front part of the vocal tract, and so alter vowel formant frequencies. Mouth and lip shapes also determine facial emotional expressions, e.g., disgust is typically expressed with a distinctive lip and mouth configuration (i.e., closed mouth, pulled back lip corners). This overlap of speech and emotion gestures suggests that expressive speech will have different vowel formant frequencies from neutral speech. This study tested this hypothesis by comparing vowels produced in neutral versus disgust expressions. We used our database of five female native Cantonese talkers each uttering 50 CHINT sentences in both a neutral tone of voice and in disgust to examine five vowels ([ɐ], [ɛː], [iː], [ɔː], [uː]). Mean fundamental frequency (F0) and the first two formants (F1 and F2) were calculated and analysed using mixed effects logistic regression. The results showed that the disgust vowels showed a significant reduction in either or both formant values (depending on vowel type) compared to neutral. We discuss the results in terms of how vowel synthesis could be used to alter the recognition of the sound of disgust.
Index Terms: emotional speech production, acoustic analysis, disgust

1. Introduction

Of the six basic emotion types, disgust seems to be the most elusive to define in terms of an acoustic profile. For example, there are conflicting results regarding how F0 is used in the vocal expression of disgust with disgust reported to have a rising pitch contour by some studies and while others found a falling one (see [1]). So while fundamental frequency is one of the most salient carriers of emotion information in the voice, it is not effective in flagging disgust. In this paper we investigated if other acoustic properties, namely the first two formant frequencies can more reliably cue disgust. This is based on the observation that the configuration of the facial muscles (especially around the mouth region) during the expression of disgust may have a direct impact on the vowel formant frequencies.

The emblematic facial expression of disgust consists of nose wrinkling; and the retraction and tightening of the lips [2, 3]. It is widely accepted that for disgust, such gestures are an adaptive response to help prevent contaminants from entering our body. Given that the facial expression of disgust basically involve the lower half of the face, this expression is likely to impose constraints on how simultaneous speech may be produced. Indeed, it has been shown by using motion capture, that the expression of disgust can result in a lowering of the larynx and cause greater retraction of the lips compared to speech spoken with a neutral expression (herewith, neutral speech) [4]. There is however, a lack of studies on tongue position during the production of spoken expressions of disgust, so we can only speculate that since the lips need to be open to produce speech, the tongue may be retracted further back into the oral cavity to act as a secondary barrier to prevent the ingestion of contaminants. The retraction of the lips and tongue will inevitably alter the shape and length of the oral cavity thereby affecting formant frequencies/articulation of vowels.

The retraction of the lips that occurs when expressing disgust would tend to alter the shape and length of the oral cavity and thereby affect the formant frequencies of articulated vowels. In regard to how a face gesture can influence formant frequencies, a parallel can be drawn with arguments made about smiled speech. Here, it has been demonstrated that the retraction of the lips during smiled speech can lead to significantly lower F1 values citeTartter1980. Based on this, we would expect to see a general reduction in F1 during the spoken expression of disgust when compared to neutral speech.

The effect that the expression of disgust may have on speech production likely extends beyond the configuration of the lips. It has been suggested that tongue position is also a major component of disgust with it being suggested that the extruded tongue may be a reflex related to expelling a contaminant [2]. However within the context of spoken expressions of disgust in general, it is not clear what the prototypical tongue position would be, i.e., in general, there is the lack of studies on tongue position during production of expressive speech. Our speculation is that the tongue will be retracted further back into the oral cavity (acting as a secondary barrier to any contaminant since in speech, the lips themselves are not kept shut). If the tongue is retracted, then F2 may be lowered (i.e., somewhat similar to the production of back vowels).

The current study tested whether F1 and F2 would be lowered in the spoken expression of disgust in comparison to neutral speech. Also of interest is whether any change in formant frequencies will be the same across the vowel types. For instance, given that the vowel [i] is already produced with a low F1, it may be that the expression of disgust will further lower this value. Alternatively, since F1 may be already at floor, disgust may be most efficiently conveyed by changing only F2. In this regard, the Vowel [u] may be the most interesting since it has low F1 and F2 values in neutral speech.

We used our Cantonese auditory-visual expressive speech database to specifically examine the acoustic correlates of disgust. In this database, there are 50 sentences produced with disgust and a matching 50 produced with a neutral tone of voice. These sentences have been spoken by five female speakers yielding a total of 500 utterances in all. Mean f0, F1 and F2 values were extracted from all the vowels, but due to length constraints, we will only describe five ([ɐ], [ɛː], [iː], [ɔː], [uː]) of these. These five were selected as they cover a large region of the vowel space. The f0 measure is included for comparison purposes.

http://dx.doi.org/10.21437/Interspeech.2016-1463

2. Methods

2.1. Speakers

The speakers consisted of five females (average age of 28.5 years, sd = 2.1) who were born and raised in Hong Kong and Cantonese is their native tongue.

2.2. Speech Materials

Fifty semantically neutral sentences were chosen from the Cantonese Hearing In Noise (CHINT) sentences list [5] on the basis that they had a good spread of different tones at the initial and final position in the sentences. These sentences were produced in different emotional tones of voice (i.e., the six basic emotion expressions) by five female speakers. Only disgust and neutral recordings of five female speakers were used in this study, resulting in a total of 500 utterances.

2.2.1. Recording Setup

While only the audio recordings of disgust and neutral are used in this paper, the entire recording procedure including video recording is reported for completeness.

Each speaker was seated in front of a 20.1" LCD video monitor (Diamond Digital DV201B) that is used to present the stimulus sentences to the speaker. Directly above the monitor was a video camera (Sony NXCAM HXR-NX30p) where the speaker was requested to fixate at prior to expressing the selected sentences. The videos were recorded at 1920 x 1080 full HD resolution at 50 fps. To capture the speakers utterances a microphone (AT 4033a Transformerless Capacitor Studio Microphone) was placed about 20 cm away from the speakers 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.

2.3. Procedure

Speakers read a scenario that was designed to elicit the target emotion. They were then given three practice trials before the actual recording commenced. These practice trials allowed the participant to familiarise themselves with the task of expressing themselves in the particular emotion type and enabled the experimenter to ensure that the audio quality was appropriate. When the speaker was ready, each stimulus sentence was displayed one at a time in a random order and the speakers produced the utterances at their own pace. The speaker was required to produce each sentence as naturally as possible and to do so with communicative intent (i.e., try to convey their feelings to the observer. Feedback was provided via the screen if sentences had to be repeated (e.g., the speaker misread the sentence or did not fixate on the camera while producing the expressions). It is important to note that other than this feedback, the experimenter did not interfere or comment on the production of the expressions.

2.4. Analysis

We used EasyAlign [6], a force alignment tool implemented in Praat [7] that is freely available to provide the initial segmentation of the auditory renditions. For this, the Cantonese sentences were first transcribed into Jyutping and then into the closest SAMPA approximation of Spanish using a custom Matlab script [8]. All of the aligned textgrids were manually checked and corrected by the first author in Praat [7].

From the 500 utterances, there were a total of 510 instances of [ɐ], 202 of [ɛː], 714 of [iː], 350 of [ɔː] and 182 of [ʊː] in both disgust and neutral. The formant values were then extracted using Praat and screened for outliers using the mvoutlier package [9] in R [10]. Outliers (total of 22 vowels) were generally due to durations that were too short or to lip smacking. These were replaced using mean substitution, i.e. the mean of the vowel, produced by the speaker in whichever expression the outlier was from. Missing values (3 vowels, due to the omission of words in the recording procedure) were similarly treated.

3. Results

Figure 1 shows the F1 and F2 values of all five vowels when produced in a disgust and neutral tone of voice. As can be seen, there is a general shift in the vowel space for disgust compared to neutral. All of the vowels had lower F1 when produced in a disgust tone of voice. As for F2, all of the vowels except [uː] and [oː] were lower in disgust.

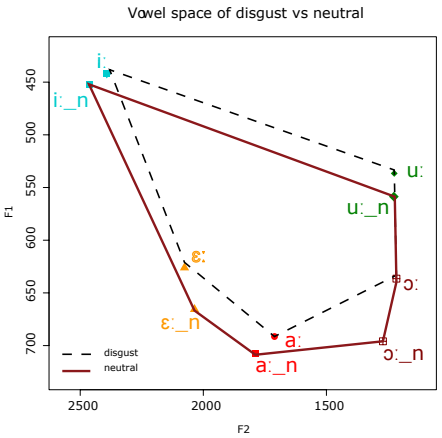

Figure 1: *Comparison of the vowel space for disgust and neutral utterances averaged across all speakers and vowel tokens. The solid line on the plot and the vowels appended with '_n' represent neutral.*

A linear discriminant analysis was conducted for each vowel to capture the pattern of how F0, F1 and F2 vary depending on emotional tone of voice. These were conducted using the MASS package [11] in R [10] and summarised in Table 1. We then examined the contribution of each acoustic measure (predictor) by conducting separate mixed effects logistic regression models for each vowel using the lme4 package [12]. The estimates were based on maximum likelihood using the Laplace Approximation method. Speaker, sentence and the position of the vowel in the sentence (1-10 depending on the position of the word where the vowel originated from) were entered as random effects and mean F0, F1 and F2 as fixed effects. It should be noted that we did not take into account the tone type of vowels as the data used in the current examination are not large enough.

The predictors were entered into the model one at a time to evaluate how the variance accounted for changed as a function of said predictor with significance determined by chi-square probability. The results were summarised in Table 2. These

Table 1: *Group means and coefficients of discriminant function*

Vowel	Measure	Mean disgust	Mean neutral	Coefficients
[ɐ]	F0	210.37	209.01	−0.0051
	F1	691.33	707.33	0.0036
	F2	1710.97	1788.58	0.0041
[ɛː]	F0	203.93	203.19	0.0060
	F1	626.07	665.47	0.0106
	F2	2076.37	2032.92	−0.0006
[iː]	F0	214.78	209.43	−0.1212
	F1	441.48	451.84	0.0093
	F2	2394.41	2464.33	0.0029
[ɔː]	F0	216.21	207.92	−0.0076
	F1	636.45	695.93	0.0098
	F2	1215.72	1270.52	0.0010
[uː]	F0	214.20	215.41	0.0072
	F1	536.66	558.69	0.0109
	F2	1224.59	1226.12	−0.0008

along with the results of the linear discriminant analysis will be explored later.

For all vowels, the mixed effects models showed that the random effects only accounted for a very small percentage of the variance (less than 1%), This was likely due to the design of our study where speaker variance and other factors were controlled by comparing the same sentences produced by the same speakers in different tones of voice. Below are the detailed results of each vowel.

Table 2: *Proportion of variance explained by the measures of mean F0, F1 and F2 as a function of vowel type.*

Vowel	Measure	Var. exp.	Sig. (χ^2)	Direction
[ɐ]	F0	0.17	ns	-
	F1	3.17	ns	-
	F2	15.23	$p < .001$	decrease
[ɛː]	F0	0.00	ns	-
	F1	10.52	$p < .01$	decrease
	F2	.07	ns	-
[iː]	F0	2.87	ns	-
	F1	4.85	$p < .05$	decrease
	F2	10.52	$p < .01$	decrease
[ɔː]	F0	4.95	$p < .05$	increase
	F1	80.24	$p < .001$	decrease
	F2	8.43	$p < .05$	decrease
[uː]	F0	.04	ns	-
	F1	5.16	$p < .05$	decrease
	F2	0.23	ns	-

3.1. The vowel [ɐ]

The linear discriminant analysis revealed that the production of [ɐ] in disgust had higher f0 values ($b = -0.0051$), and lower F1 and F2 values ($b = 0.0036$ and $b = 0.0041$) (see Table 1. The mixed effects logistic regression however showed that f0 and F1 were not useful in discriminating disgust from neutral sounds (proportion of variance accounted for, 0.17 and 3.17. Note that for the coefficients from the mixed effects model, a positive value indicated that the value was larger in neutral speech compared to disgust while the coefficients from the discriminant analysis showed the opposite pattern, i.e., the value was larger for disgust compared to neutral.

The model significantly improved when F2 was added, $b = .0016$, mean (disgust, 1710.97, neutral, 1788.58), $p < .001$.,. The odds ratio showed that as F2 decrease by one more unit, the change in odds that an utterance was produced in a disgust tone of voice was 1.0016 (confidence interval, 5%, 1.0008 and 95%, 1.0024).

3.2. The vowel [ɛː]

The linear discriminant analysis showed that for this vowel, disgust can be differentiated from neutral by a decrease in f0 and F1 ($b = 0.0060$ and $b = 0.0106$) and an increase in F2 ($b = -.0006$). The mixed effects logistic regression however showed that F0 and F2 were not useful in discriminating disgust from neutral, with the low proportion of variance accounted for (0.00 and 0.07).

Only F1 significantly improved the model, 8.52, $b = 0.004$, mean (disgust,626.07, neutral, 665.47) $p < .01$.. The odds ratio showed that as F1 decrease by one more unit, the change in odds that an utterance was produced in a disgust tone of voice was 1.0048 (confidence interval, 5%, 1.0014 and 95%, 1.0082).

3.3. The vowel [iː]

Disgust can be differentiated from neutral by a decrease in F1 and F2 ($b = 0.0093$ and $b = 0.0029$) and an increase in f0 ($b = -.1212$). The mixed effects logistic regression showed that only F0 was not useful in discriminating disgust from neutral, proportion of variance accounted for (0.00).

F1 and F2 significantly improved the model $b = 0.0024$, mean (disgust,441.48, neutral, 451.84) $p < .05$ and $b = 0.0047$, mean (disgust,2394.41, neutral, 2464.33) $p < .001$., respectively. The odds ratio showed that as F1 decrease by one more unit, the change in odds that an utterance was produced in a disgust tone of voice is 1.0011 (confidence interval, 2.5%, 1.0005 and 95%, 1.0016). As F2 decrease by one more unit, the change in odds that an utterance was produced in a disgust tone of voice was 1.0034 (confidence interval, 5%, 1.0012 and 95%, 1.0057).

3.4. The vowel [ɔː]

The discriminant analysis showed that this vowel was produced with decreased F1 and F2 ($b = 0.0098$ and $b = 0.0010$) with increased f0 ($b = -.0076$) in a disgust compared to neutral tone of voice. The mixed effects logistic regression showed that all the predictors were able to account for significant proportions of the variance (F0, 4.95, F1, 80.24 and F2, 8.43). Of the three predictors, F1 showed the greatest amount of variance explained. Given the large variance, we double checked for outliers and influential data points but none were to be found.

F0 significantly improved the model, $b = 0.0027$, mean (disgust,216.21, neutral,207.92), $p < .05$.. The odds ratio showed that as F0 increase by one more unit, the change in odds that an utterance was produced in a disgust tone of voice was 0.9947 (confidence interval, 5%, 0.9905 and 95%, 0.9989).

F1 significantly improved the model, $b = 0.0469$, mean (disgust,636.45, neutral,695.93), $p < .001$.. As F1 decrease

by one more unit, the change in odds that an utterance was produced in a disgust tone of voice was 1.0071 (confidence interval, $5\%, 1.0052$ and $95\%, 1.0090$). Likewise, the addition of F2 significantly improved the model, $b = 0.0024$, mean (disgust,1215.72, neutral,1270.52), $p < .05$.. As F2 decrease by one more unit, the change in odds that an utterance was produced in a disgust tone of voice was 1.0001 (confidence interval, $5\%, 0.9999$ and $95\%, 1.0014$).

3.5. The vowel [uː]

For this vowel, disgust was differentiated from neutral by an overall decrease in f0, F1 and F2 ($b = 0.0072$, $b = 0.0109$ and $b = -.0008$). The mixed effects logistic regression however showed that the addition of F1 significantly improved the fit of the model, $b = 0.0026$, mean (disgust,536.66, neutral,558.69) $p < .05$.. As F1 decrease by one more unit, the change in odds that an utterance was produced in a disgust tone of voice was 1.0026 (confidence interval, $5\%, 1.0004$ and $95\%, 1.0048$).

4. Discussion

In this study, we proposed that spoken expressions of disgust will result in different vowel formant frequencies when compared to those produced from neutral speech. To test the proposal, we compared F1 and F2 (as well as f0) of 5 vowels produced in neutral versus disgust expressions. Out of the 5 vowels examined, 4 of them (all tested vowels except [ɐ]) showed a significant decrease in F1 while only three vowels ([ɐ], [iː] and [ɔː]) showed a significant decrease in F2 when produced in a disgust tone of voice. Only one vowel, ([ɔː] showed an increase in f0. In line with our hypothesis, the measures of F1 and F2 may be more reliable than F0 is in marking disgust. However there may potentially be language effects since Cantonese as a tone language and has been shown to utilise less F0 information in vocal expressions of emotion [13, 14]). Our next step is to examine if inflections of F0 in disgust may interact with lexical tones.

From our results, the vowels produced in disgust can generally be characterised by a lowering of F1. Interestingly, the change in F1 appears to be most pronounced for rounded vowels that require lip protrusion such as [uː] and [ɔː]. We examined the video clips that were recorded together with the audio recording and found that these rounded vowels appear to be produced with less lip protrusion and rounding when produced in disgust, see Figure 2. From this, our speculation is that disgust may be most efficiently conveyed by reducing lip protrusion of these vowels so that salient acoustic changes may be produced thus assisting detection by a listener. This can be verified by examining if these vowels may be emphasised during the production of vocal expressions of disgust such that these vowels may be sustained (increase in duration when compared to other vowels in the sentence) to maximise the salience of these formant changes.

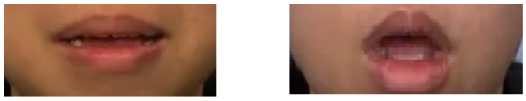

Figure 2: *The figure on the left shows the configuration of the lips when producing the vowel [ɔː] in disgust while the figure on the right shows the same vowel produced by the same speaker from the same word and from the same sentence but in neutral.*

While we were unable to measure the placement of the tongue in this study, the change in F2 suggests that the vowels were produced with place of articulation that is further back from normal suggesting that there may indeed be underlying changes in tongue position.

With regards to the implications of these results, first, from the changes in the formant frequencies of the 5 vowels, we can predict how the other vowels/diptongs may be produced. It will be interesting to examine dipthongs since they are combinations of the 5 vowels that we have examined. For example, we can predict that [ou] may be produced with lower F1 but with no changes in F2. This is currently underway.

Second, our results may shed some light as to why disgust is generally found to be one of the emotions that is hardest to recognise when presented in an auditory only condition [15, 16]. If changes in formant frequencies are the most salient acoustic marker of disgust, there is a need for a speaker to first establish a baseline for a neutral vowel space in order for a listener to be able to recognise the subtle departures in formant frequencies. This may require prolonged exposure and is in contrast with other emotion types such as sadness where fundamental frequency and speech rate are the key characteristics of the emotion.

Third, the results of this study taken together with the studies on smiled speech [17] suggest that the vocal expressions of emotions contain information pertaining to the facial expression of an emotion rather than merely to the valence of the emotion itself. So rather than looking for acoustic markers that code for specific dimensions such as arousal or valence, it may be more fruitful to search for measures that correlate with the facial expression of an emotion. Moreover emotion information is most salient in the face [18, 19] and vocal expressions of emotions are often (if not always) produced together with facial expressions.

This study is somewhat exploratory in nature given that it is the first in a series of planned experiments. Along with the other factors that we have outlined above, we plan to verify if the same pattern of results can be observed with other languages (English). If disgust serve an adaptive role and its prototypical facial expression is universal across cultures and languages, a similar change in the formant frequencies of spoken expressions of disgust should be observed. We also plan to conduct a perception study using synthesized vowels/speech to investigate if listeners may be able to discriminate disgust from neutral based on formant frequency changes alone. Moreover it is unlikely that the mean of the first two formant frequencies are the only properties that may code for disgust, it is undeniable that they play a rather large role in the expression of disgust. Certainly there is a need to further examine how pitch, duration and other acoustic properties may be used in tandem with the changes in formant frequencies.

5. Conclusion

This study showed that the facial expression of disgust affect the formant values. This was most likely due to the retraction of the lips during the expression of disgust which had the effect of lowering F1 and F2. This finding suggests that these measures may stand as possible acoustic markers of the expression of disgust.

6. References

[1] K. R. Scherer, "Vocal correlates of emotional arousal and affective disturbance." pp. 165–197, 1989.

[2] P. Rozin, J. Haidt, and C. McCauley, "Disgust," *Handbook of Emotions*, pp. 757–776, 2008. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2008-14016-001&site=ehost-live

[3] 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," *Journal of Personality and Social Psychology*, vol. 53, no. 4, pp. 712–717, 1987.

[4] G. Bailly, A. Bégault, F. Elisei, P. Badin, and C. G. Universities, "Speaking with smile or disgust : data and models," *Interspeech*, pp. 111–114, 2008.

[5] 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

[6] J. P. Goldman, "EasyAlign: An automatic phonetic alignment tool under Praat," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3233–3236, 2011.

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

[8] "Matlab R2013a."

[9] P. Filzmoser and M. Gschwandtner, *mvoutlier: Multivariate outlier detection based on robust methods*, 2015, r package version 2.0.6. [Online]. Available: http://CRAN.R-project.org/package=mvoutlier

[10] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org/

[11] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[12] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, pp. 1–51, 2014. [Online]. Available: http://arxiv.org/abs/1406.5823

[13] C. S. Chong, J. Kim, and C. Davis, "Exploring Acoustic Differences between Cantonese (Tonal) and English (Non - Tonal) Spoken Expressions of Emotions," *submitted*, 2015.

[14] P. a. Hallé, Y.-C. Chang, and C. T. Best, "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," *Journal of Phonetics*, vol. 32, no. 3, pp. 395–421, jul 2004. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0095447003000160

[15] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, pp. 614–36, mar 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8851745

[16] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, apr 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167639302000845

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

[18] 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.

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