Articulatory characteristics of emotional utterances in spoken English
Donna Erickson,1 Arthur Abramson,2 Kikuo Maekawa,3 Tokihiko Kaburagi†

1Gifu City Women’s College, Gifu, Japan; 2Haskins Laboratories, New Haven, CT, U.S.A.; 3National Language Research Institute, Tokyo, Japan; 4 Kyushu Inst. Design, Fukuoka, Japan

ABSTRACT
Acoustic and articulatory properties of emotional utterances in English were examined using articulatory (EMA) recordings of speech elicited from two speakers of American English. The speakers produced 10 to 12 repetitions of the sentence “That’s wonderful,” using several different intonational patterns and types of paralinguistic information. Perception tests showed that listeners could perceive the emotions intended by the speakers. Furthermore, F0, formant frequencies, jaw and tongue dorsum position changed as a function of the particular emotion. Initial analysis suggests that the emotion “anger” may involve more jaw lowering, “suspicion,” a raising of the tongue, and “admiration,” a lowering of the tongue.

1. INTRODUCTION

Previous work has shown that emotion changes the acoustic and articulatory characteristics of speech. It has been reported that intonation contours change as a function of the emotional content of the utterance, as do duration, voice quality, loudness, etc [e.g., 1, 2]. As for intonational contours, Erickson [3], however, reported that similar F0 contours could express different emotions if other characteristics, such as voice quality, changed. This suggests that speakers can change the emotional (paralinguistic) content of an utterance independently of the intonational (linguistic) parameters.

As for articulation, Erickson et al. [4] reported an increase in jaw opening with increased irritation. Maekawa et al. [5,6] showed that the tongue dorsum position and formant frequencies were affected by the paralinguistic (emotional) content of the utterance. Specifically, those that were spoken with “suspicion” were uttered with a more forward tongue, while those spoken with “admiration” were produced with a more backed tongue.

This study examined some acoustic and articulatory characteristics of emotional utterances in spoken English. The intonational patterns and emotion types were based on work done earlier for American English [3] and also for Japanese [5].

These questions were asked: (1) Can listeners perceive emotions as intended by the speakers? (2) What are some intonational contours that can occur with different emotions? (3) Does articulation change with the different emotions, and if so, in what way? (4) How do formant frequencies change? (5) How does this compare with the findings for Japanese emotional utterances?

2. METHODS

In order to examine the articulatory characteristics of emotional utterances in spoken English, articulatory EMA (Electromagnetic Articulograph) recordings of speech were elicited from two speakers of American English (one male, one female) at the NTT Laboratories, Atsugi, Japan (courtesy of M. Honda) [7]. The speakers (the first two authors) produced 10 to 12 repetitions of the sentence, “That’s wonderful,” using several different intonational patterns (as transcribed with the ToBI system [8]) and several types of paralinguistic information. The female speaker (Speaker 1) produced 10 repetitions of this sentence using 7 different intonation contours with several different emotions. The male speaker (Speaker 2), recorded 8 months later, produced 12 repetitions of a similar set of intonational contours with “That’s wonderful.” In addition, Speaker 2 produced the same ten intonation contours as “neutral” utterances. Before recording, the speakers rehearsed the utterances several times until they felt comfortable with the intended contours and emotions.

The EMA recordings of articulatory motions were made by attaching receiver coils to the surface of the jaw, upper lips, lower lip, and tongue of each subject and measuring the position of the coil at a sampling frequency of 250 Hz. Tongue motion was measured at three and four points for the male and female subjects, respectively. For this study, we examined the jaw and tongue dorsum (coil 5) for each of the subjects.

We report here on a subset of the data for Speaker 1, three emotions (admiration, suspicion and anger) on three different intonational patterns (L+H L- L%, L*H- H%, and H*H=L-L%, respectively); for Speaker 2, four emotions (admiration, suspicion, anger, and disappointment) on five different intonational patterns (L+H*L- L%, L*H- H%, L*H H- L* H- H%, H*H=L-L%, and L*L=L-L%; the second and third patterns are both for “suspicion”). Notice that the first intonation contour for “suspicion” and that for “anger” are the same for both speakers; for “admiration,” the intonational contours are different. For Speaker 2, due to a problem with the coils staying glued to the tongue, it was possible to analyze only 6 of the 12 repetitions. For Speaker 1, only 5 of the 10 repetitions were analyzed.

Articulatory measurements of the jaw and tongue x-y positions were made at the time of maximum jaw opening during the vowel portion in the word that’s and the first syllable of the word wonderful for each of the utterances. Formant-frequency measurements were also made at the time of maximum jaw opening. The articulatory and acoustic measurements were made with a MATLAB-based program written by Jianwu Dang (ATR, Kyoto, Japan). F0 contours, used for the ToBI analysis, were produced with the ESPS WAVES software.

3. PERCEPTION TESTS

To validate the association of our intonational patterns with the intended emotional categories, we tested them perceptually with native speakers of American English. For each speaker, two tokens of each utterance were used, yielding 20 responses to
tone, and since the pitch tracker was not able to track its low glottal rate, the F0 contour is incomplete during the L*.

With regard to “anger,” the ToBI transcriptions are the same, but for Speaker 1, the F0 tends to be high, with no downturn on the second H*. For Speaker 2, the F0 tends to be somewhat lower in the speaker’s F0 range, with considerable downturn on the second H*. The mean values of F0 (Hz), along with formants (Hz), jaw, and tongue x-y positions (mm) measured at the time of maximum jaw opening during the vowels of that’s and the first syllable of the word wonderful are shown in Table 3.

### Table 3. Mean values of F0, formants, jaw and tongue.

<table>
<thead>
<tr>
<th></th>
<th>Speaker 1</th>
<th>Speaker 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>valu</td>
<td>sus</td>
</tr>
<tr>
<td>J1</td>
<td>360</td>
<td>236</td>
</tr>
<tr>
<td>Jx</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>Jy</td>
<td>127</td>
<td>129</td>
</tr>
<tr>
<td>T3x</td>
<td>104</td>
<td>106</td>
</tr>
<tr>
<td>T3y</td>
<td>148</td>
<td>149</td>
</tr>
<tr>
<td>F0</td>
<td>153</td>
<td>150</td>
</tr>
<tr>
<td>F1</td>
<td>270</td>
<td>331</td>
</tr>
<tr>
<td>F2</td>
<td>1130</td>
<td>1403</td>
</tr>
<tr>
<td>Jx</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>Jy</td>
<td>126</td>
<td>128</td>
</tr>
<tr>
<td>T3x</td>
<td>109</td>
<td>107</td>
</tr>
<tr>
<td>T3y</td>
<td>145</td>
<td>151</td>
</tr>
</tbody>
</table>

Scatter plots of F1 vs. F2, jaw x vs. jaw y, and tongue dorsum x vs. y for the two vowels for Speaker 1 are shown in Figure 2, and for Speaker 3, in Figure 3. The F1–F2 axes are reversed to make the acoustic space match the articulator space for the jaw and tongue scatter plots, as if the speaker were facing the left side of the page. For Speaker 1, even though the vowels differ in terms of formant-frequency values and jaw/tongue positions, we see similar acoustic and articulator patterns for both vowels with the jaw and tongue lowest for “anger” and highest for “suspicion,” corresponding to raised F1/F2 for “anger” and lowered F1/F2 for “suspicion.” For Speaker 2, for both vowels, “suspicion” always has a higher jaw and tongue position, relative to both “anger” and “admiration,” for both vowels. However, for “admiration” there is a difference in articulation for the two words. For won there is a much lower jaw than for that, even though [æ] is generally a lower vowel than [æ], and we would expect to see more jaw opening with the more open vowel e.g., [9]. The F0 for “admiration” on won is extremely high (236 Hz), but also notice that the jaw position is more forward. Forward jaw position may help expedite the production of the combination of high F0 and very low jaw [10].

MANOVA showed that the means of each of the measured acoustic and articulatory values were significantly different (p<.01) as a function of the emotion. This finding corresponds well to the perception results, which showed that the emotions were well-perceived by listeners. It is interesting that listeners tended to show some confusion between “suspicion/disbelief” and “admiration.” In the acoustic and articulatory data displayed in the scatter plots for Speaker 1, there seems to be some overlap...
in formant frequencies and jaw and tongue dorsum position in the production of these two emotions. From this point of view, it is not surprising that “suspicion/disbelief” utterances were sometimes heard as “admiration.” As for Speaker 2, there is virtually no overlap between these two emotions in terms of formant frequencies; however, for the word that’s, there is considerable overlap of jaw xy coordinates. Investigation of acoustic and articulatory characteristics as they correlate with the perceptual confusion matrix will be explored more closely as we run additional perception tests presenting listeners with single words, rather than the entire sentence.

5. SUMMARY

The intended emotions of the two speakers were well-perceived by listeners. Formants and articulatory settings were affected by the emotional content of the speaker. Different intonational patterns can be used for similar emotions. If we compare these two American English speakers with the one Japanese speaker [5,6], it seems that “suspicion” and “admiration” in both English and Japanese involve changes in tongue-dorsum positions—either a raising or fronting of the tongue for “suspicion” and lowering or backing for “admiration.” However, for the Japanese speaker, assignment of emotion involved the same articulatory setting throughout the utterance, even for the consonants; for one of the American English speakers, the articulatory setting was different for the first and second vowels for the production of “admiration.”

Future work is underway to examine the relation between F0 height, intonational patterns, and articulation, as well as the durational and voice quality characteristics of these utterances. Also, we wish to explore interspeaker differences in production of emotion, specifically, changes in laryngeal source configuration vs. changes in supraglottal configuration. In the preliminary study here, it seemed that Speaker 2 tended to use changes in voice source characteristics (e.g., glottal fry for “suspicion”), whereas Speaker 1 tended to use changes in supraglottal articulation to signal emotion.

6. REFERENCES


Figure 1. Intonation contours and ToBI transcriptions for Speaker 1 (left column) and Speaker 2 (right column) for “That’s wonderful” spoken on “admiration” (top panels), “suspicion/disbelief” (middle panels), and “anger” (bottom panels).
Figure 2. Speaker 1. Top row shows formant, tongue dorsum and jaw measurements at point of maximum jaw opening during [ae]; bottom row shows this for [ ]. “A” indicates “admiration;” “S,” “suspicion;”; and “N,” “anger”.

Figure 3. Speaker 2. Top row shows formant, tongue dorsum and jaw measurements at point of maximum jaw opening during [ae]; bottom row shows this for [ ]. “A” indicates “admiration;” “S,” “suspicion;”; and “N,” “anger”.