Identifying Emotion in Speech Prosody Using Acoustical Cues of Harmony

Takashi Fujisawa & Norman D. Cook

Department of Informatics
Kansai University, Osaka, Japan
fujisawa_jr@hotmail.com, cook@res.kutc.kansai-u.ac.jp

Abstract
We have studied the prosody of emotional speech using a psychoacoustical model of musical harmony (designed to explain the basic facts of the perception of pitch combinations: interval consonance/dissonance and chordal harmony/tension). For any voiced utterance, the model provides 3 quasi-musical measures: dissonance, tension, and harmonic “modality” of the pitches used. Modality is the most interesting, as it relates to the major and minor modes of traditional harmony theory and their characteristic positive and negative affect. In a study of emotional speech using 216 utterances, factor analysis showed that these measures are distinct from those obtained from basic statistics on the fundamental frequency of the voice (mean F0, range, rate of change, etc.). Moreover, there was a significant correlation between the major/minor modality measure and the positive/ negative affect of the utterance. We argue that, in addition to the traditional acoustical measures, a measure of multiple-pitch combinations, i.e., harmony, is essential for determining the affective character of the tone of voice in speech.

1. Introduction

The central paradox in the study of emotional prosody is the fact that affective information carried in the fundamental frequency (F0) of the voice, but various acoustic measures (mean F0, range, rate of change, formant structure, etc.) do not allow linguists to distinguish between anger and joy nor between contentment and sadness. Despite decades of research, it has remained uncertain what features of the voice are relevant to the expression and understanding of emotional speech. As a consequence, neither acoustical measures of the voice nor various qualitative descriptions of pitch contours suffice to identify emotions that normal listeners readily identify. What then is missing from the analysis of pitch prosody?

2. Remaining Problems in the Study of Emotional Speech

In a review of the diverse literature on vocal affect expression, Scherer [1] has summarized the principal factors that convey emotion, as shown in Table 1. Voices expressing "joy", "anger" and "fear" characteristically show high pitch levels, wide pitch ranges, fast tempos and so on. In contrast, “sadness”, "boredom" and "indifference" are expressed with low pitch levels, narrow pitch ranges, slow tempos and so on.

However, these findings lead to the negative conclusion that the acoustical parameters measured in previous studies cannot distinguish between markedly different emotional states. Using these parameters, we can only identify two groups of emotions. One is the high frequency/amplitude/ tempo group corresponding to a high degree of physiological arousal, and another is the low frequency/amplitude/tempo group corresponding to a low degree of physiological arousal [1]. From the viewpoint of the circumplex model of emotional dimensions [2], it is clear that identification of the dimension corresponding to “valence” relative to “arousal” remains uncertain. For example, we can distinguish easily between anger and sadness by using these parameters, but cannot distinguish between anger and happiness.

<table>
<thead>
<tr>
<th>Table 1: Summary of Results on Vocal Indicators of Emotional States [1].</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Emotion</td>
</tr>
<tr>
<td>Level</td>
</tr>
<tr>
<td>Range</td>
</tr>
<tr>
<td>Loudness</td>
</tr>
<tr>
<td>Tempo</td>
</tr>
</tbody>
</table>

Scherer, Johnstone, & Klasmeyer [3] state that this is one of the unresolved problems in the study of emotional prosody: how to define "intonational patterns" and extract the relevant features. In the present study, we propose new acoustical parameters which can be extracted from pitch contours, and evaluated quantitatively. These parameters, as discussed below, are essentially the same acoustical cues that are used to express emotion in music. Since the intonation of speech and the melody of music are both pitch phenomena, it may be that the same acoustical cues are used in both realms to express emotions.

3. Acoustical Cues of Emotional Expression in Music

In an attempt to clarify acoustical parameters used in the emotional expression of music, Gabrielson & Justlin did a meta-analysis of published empirical research on emotional expression in music [4]. They concluded that there is a consistent tendency in the use of certain acoustical features to express each emotional category.

Clear tendencies were found for categories such as happiness, joy and gaiety, as well as categories such as sadness and gloom. The acoustical features of music expressing happiness are consonance, major mode, high pitch level and wide range of pitch, high intensity level and quick tempo. On the contrary, sad music is most often expressed...
through dissonance, the minor mode, low pitch levels, a narrow range of pitch, low intensity levels and slow tempo. In other emotional categories, for example, anger, high intensity, fast tempo and high pitch levels are used, sometimes in combination with dissonance and the minor mode.

### Table 1

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Dissonance (1)</th>
<th>Consonance (4)</th>
<th>Mode: Major (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>Pitch level: High (3)</td>
<td>Loudness: High (13)</td>
<td>Tempo: Fast (13)</td>
</tr>
<tr>
<td>Sad</td>
<td>Pitch level: Low (8)</td>
<td>Loudness: Soft (12)</td>
<td>Tempo: Slow (24)</td>
</tr>
<tr>
<td>Relaxed</td>
<td>Pitch level: High (1)</td>
<td>Loudness: Low (1)</td>
<td>Tempo: Slow (24)</td>
</tr>
</tbody>
</table>

Figure 1: Mapping of the acoustical features of musical affect. The number following the acoustical feature expresses the number of literature studies indicating this feature [4].

By mapping the acoustical features used to express each emotional category onto the dimensions of the circumplex model, interesting results were obtained. Four emotional categories (happy, sad, angry, relaxed) were selected arbitrarily, and the acoustical features in each category were mapped. As shown in Figure 1, two conclusions can be drawn from the relationship between the emotional dimensions and each acoustical feature.

1) High/low pitch level, wide/narrow pitch range, loud/soft intensity and fast/slow tempo is correspond to the "arousal" dimension.

2) In contrast, consonance/dissonance and the major/minor mode correspond to the positive-negative affect dimension of the emotion. These acoustical indices correspond to the "valence" dimension.

Notably, the outstanding acoustical parameters in music were found to be dissonance/consonance and major/minor modality. The fact that linguists have not been able to differentiate emotions corresponding to the "valence" dimension in voice expression may well be due to the fact that musical features, especially "harmoniousness", has not been considered. In order to evaluate the "harmony" of speech, two technical problems must be solved:

1) "Consonance" and "harmoniousness" must be defined as quantitative parameters based on psychophysical acoustics, not as empirical categories from music theory.

2) A technique for determining the discrete "pitches" that are the basis for the harmonious characteristics of the voice (for calculation of the "consonance/dissonance" and "major/minor" modality) is required. Discrete pitches must be extracted from the continuously changing pitch contour.

### 4. Model of Harmony Perception

We have found that harmoniousness and major/minor modality can be expressed quantitatively provided that both 2-tone effects (the consonance/dissonance of intervals) and 3-tone effects (the sonority/tension of 3-tone chords) are considered as separate factors in the psychophysical model. Two-tone effects have already been adequately studied and interval perception has been successfully modeled [5-7].

Three-tone structures immediately introduce the full complexity of traditional harmony theory, but already in 1956 Meyer had identified the central issue of chordal "tension" as being a consequence of intervals of equal magnitude: "intervallic equivalence" [8]. That is, if a 3-tone chord contains two intervals of equivalent magnitude (e.g., the two 4-semitone intervals of an augmented chord), then it will have an inherent "tension". The stability/instability of any number of tones can therefore be calculated using an algorithm to compute the dissonance, $D$, of all tone pairs [Eq. 1], and the tension, $T$, of all tone triplets [Eq. 2].

$$D = \min\text{Amp} \ast e^{*} \left( \exp(a \ast x) - \exp(b \ast x) \right)$$

$$T = \min\text{Amp} \ast \exp(- (x_{1} - x_{2}) / d)^{2}$$

where $\min\text{Amp}$ indicates the amplitude of the pitch with the smallest amplitude, $a$, $x_{1}$ and $x_{2}$ are interval sizes (in semitones) and $a-d$ are constants (1.20, 4.00, 3.53, 0.60, respectively) chosen to produce the known (experimentally measured) relative sonority of the triads. The theoretical dissonance and tension curves are shown in Figure 1A and B. In practice, computations must be made for every pair and every triplet of tones, including the overtones (with suitable adjustments for the weaker amplitude of the higher harmonics) (the algorithm in C is available at: www.res.kutc.kansai-u.ac.jp/~cook).

Interestingly, if tension is taken as the most salient aspect of the perception of 3-tone chords, then there are two (and only two) directions in which tone combinations can move from a state of tension toward one of resolution: an increase or decrease in auditory frequency of any of the pitches of the chord. Major/minor modality, $M$, can therefore be defined in terms of the direction of pitch movement from a state of harmonic tension: the relative size of the two intervals, $x_{1}$ and $x_{2}$, in a three-tone chord [Eq. 3].

$$M = f \ast \min\text{Amp} \ast (x_{1} - x_{2}) \ast \exp(-(x_{1} - x_{2})^{2} / 2)$$

where $f$ is a constant (1.65), the intervals, $x_{1}$ and $x_{2}$, are defined in semitones; $x_{1}$ is the lower interval and $x_{2}$ is the higher interval. The modality curve is shown in Figure 1C.

The meaning of the three curves in Figure 2 can be explained simply as follows: (A) small intervals (~0.5-1.0 semitones) give high dissonance values, whereas very small or large intervals give low dissonance values. (B) Triads containing 2 equivalent intervals (a difference of intervals ~0.0) give high tension values; chords with unequal intervals (a difference of intervals ~4.10) have low tension values. (C) When the lower of the two intervals in a triad is larger than the upper interval, the modality score is positive (major-like); when the lower interval is smaller, the modality score is negative (minor-like). The simple curves in Fig. 2 become quite complex as the cumulative effects of the upper partials are brought into consideration, but the empirical findings on the relative stability of the triads and the characteristic
modality of major and minor chords (in various inversions) are reliably reproduced (see refs. [10-12]).

Figure 2: The three main factors contributing to the perception of harmony. (A) shows the dissonance factor, computed for every pair of tones and their upper partials. (B) shows the tension factor, computed for every tone triplet. (C) shows the modality factor that gives a positive (major chord-like) value or a negative (minor chord-like) value for every tone triplet. It is noteworthy that all three curves become considerably more complex when the effects of upper partials are included in the calculations, but the theoretical results reproduce the empirical findings (Table 1) (refs. [10-12]).

5. Identifying “dominant pitches”

The psychoacoustic model outlined above was designed specifically to reproduce the experimental results concerning the perception of diatonic chords [9]. Its significance for harmony theory has been discussed elsewhere [10-12], but we have recently applied these same formula to the pitch phenomena of emotional speech. In the first experiment [13-17], significantly different (major/minor) modality measures were obtained for happy and sad sentences.

We recorded “emotional” sentences, read aloud by 18 undergraduate subjects (13 male) for acoustical analysis. The sentences described typical emotional events, such as a grandparent dying or winning money in a lottery, and the subjects were instructed to read them with empathy. On the basis of the semantic content of the sentences, four were intended as having positive affect (joy, satisfaction or pleasantness), four had negative affect (sadness, anger or unpleasantness) and four were designed to express ambivalence (uncertainty, tension or anxiety) with regard to affect. Each of the 18 subjects read all 12 sentences aloud, giving a total of 216 utterances for analysis. They were allowed three trials per sentence and each subject chose the utterance that they felt best conveyed the intended emotion.

Because the affective quality of such utterances varied widely among subjects, the positive-negative valence of the 216 sentences was evaluated in a separate experiment employing a different set of 24 undergraduates. For evaluation, each utterance was converted into an unintelligible humming sequence (using the Analyze-Convert functions in Praat [18]), played through headphones at a comfortable volume adjusted by the subjects, and scored by six subjects per utterance on a 6-point scale of positive to negative affect. In this manner, the utterances perceived as affectively positive, negative or ambivalent were scored on their prosodic content alone, regardless of the speaker’s original intention.

Figure 3: A pitch histogram and its reconstruction using the cluster algorithm [19]. The fitting of Gaussian clusters (thick lines) to the raw data (thin lines) is attempted with 1-12 Gaussian curves and the “best fit” is selected on the basis of a maximum entropy calculation [20]. The modality (etc.) of the pitches in the utterance can then be calculated using the 5 pitches indicated by the 5 peaks (in this example, at 81, 108, 139, 158 and 170 Hz).

As a result of analyzing a confusion matrix consisting of the speaker’s intended affect (positive, negative or ambivalent) and the perceived affect in the utterance evaluation experiment, it was found that the positive-negative polarity of the intended affect was generally perceived by the listeners, but the intended anxiety or tension of the ambivalent sentences was not. As a consequence, further discussion of the results is made solely in terms of the perceived positive/negative affect of the utterances.

Pitch F0 was calculated at 1 millisecond intervals, giving 500-1000 pitch values per utterance. Those data were then
used as input to a “cluster” algorithm [19] that calculates a best fit between the raw data and the summation of 1-12 Gaussian clusters (radial basis functions). As shown in Fig. 3, the “clusters” simplify the raw pitch data and thus provide a small number of dominant pitches per utterance. The number of clusters per utterance is determined automatically by a maximum entropy technique [21] – normally giving 3-8 dominant pitches per utterance. Each cluster has variable position and width along the frequency axis, and variable intensity (height). The clusters are the material on which a musical analysis was done using Eqs. (1)-(3) (details of the maximum entropy technique [21] – normally giving 3-8 small number of dominant pitches per utterance. The number of “clusters” simplify the raw pitch data and thus provide a Gaussian clusters (radial basis functions). As shown in Fig. 3, the best fit between the raw data and the summation of 1-12 speech utterances of the two types of perceived affect were exerted on speech, and vice versa, continues to be debated, but quantitative techniques to test various views empirically have been lacking. The psychoacoustical model outlined in Section 2 may therefore prove useful in providing objective measures of harmony for both speech and music. The model was in fact designed to account for the most well-established harmonic phenomena of traditional Western music, but it can be used in other contexts (other musical traditions and in non-scalar pitch phenomena such as speech) because all measures are concerned with relative distances among pitches, not absolute intervals.

Acknowledgement: Supported by the “Research for the Future Program,” administered by the Japan Society for the Promotion of Science (Project No. JSPS-RFTF99P01401), Aspects of this research have previously been published in refs. [10-17].

6. Conclusion

The relationship between the quasi-musical changes in the F0 of the voice in speech and the discrete changes in the sustained tones of most musical melodies has been discussed for centuries, if not millennia. The influence that music has exerted on speech, and vice versa, continues to be debated, but quantitative techniques to test various views empirically have been lacking. The psychoacoustical model outlined in Section 2 may therefore prove useful in providing objective measures of harmony for both speech and music. The model was in fact designed to account for the most well-established harmonic phenomena of traditional Western music, but it can be used in other contexts (other musical traditions and in non-scalar pitch phenomena such as speech) because all measures are concerned with relative distances among pitches, not absolute intervals.

Table 2: MEAN SCORES OF THE MUSICAL MEASURES IN THE 25% MOST POSITIVE AND THE 25% MOST NEGATIVE UTTERANCES

<table>
<thead>
<tr>
<th></th>
<th>Dissonance (D)</th>
<th>Tension (T)</th>
<th>Modality (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos 25%</td>
<td>0.912</td>
<td>1.608</td>
<td>0.361</td>
</tr>
<tr>
<td>Neg 25%</td>
<td>1.174</td>
<td>1.552</td>
<td>-0.067</td>
</tr>
</tbody>
</table>

The mean dissonance, tension and modality of the 100 speech utterances of the two types of perceived affect were calculated. There was greater dissonance among the clusters in the Negative Affect sentences (\(\text{mean} = 2.475, p < 0.05\)). These effects reflect the fact that negative affect is typically expressed with a smaller range of F0 [1, 3], and therefore necessarily a higher dissonance value. Significant differences in tension between the 2 conditions were not found (\(\text{mean} = 0.369, p < \text{ns}\)). The most interesting results concern modality. The utterances perceived as having positive affect showed higher modality values, indicative of greater major-like pitch substructure (\(\text{mean} = 2.887, p < 0.005\)). The prediction that the sentences with negative affect would have negative (minor-like) modality scores was not found, but the relatively lower values indicate the anticipated, less positive pitch structure. The modality scores was statistically significant, but low (\(R = 0.212, p < 0.05\)). Reasons for the limited success of the experiment are numerous and will be addressed in future work.

7. References