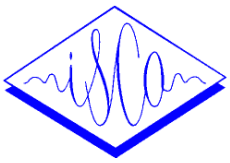

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

On the Role of Pitch Intervals in the Perception of Emotional Speech

Takashi Fujisawa, Kazuaki Takami and *Norman D. Cook*
Graduate School of Informatics, Kansai University
Takatsuki, Osaka, Japan

ABSTRACT

Evaluation of the emotional valence of speech cannot be done on the basis of first-order statistics concerning the fundamental frequency (F0), and the quantification of "pitch contours" has remained an intractable problem. We have addressed the issue of the relationship between emotion and F0 using a new technique for the extraction of the dominant tones within speech utterances and then the analysis of the interval structure. Our approach entails the summation of F0 over the entire utterance and calculation of the underlying pitch structure using an unsupervised "cluster" (radial basis function) algorithm. The technique normally results in 2-5 Gaussian "pitch clusters" per utterance that can then be evaluated in terms of their inherent dissonance and harmonic tension. We have found greater dissonance and greater harmonic tension in utterances with negative affect, relative to utterances with positive affect.

1. INTRODUCTION

In research on voice intonation, certain regularities in fundamental frequency (F0) for happy, sad, angry, fearful, etc. emotional states have been noted [1], but the ability to predict emotions solely on the basis of F0 characteristics is poor. Attempts at quantifying so-called pitch contours have thus far been inconclusive. As a consequence, there remains the paradox that normal listeners use pitch information to detect speaker emotion, but phonologists have not been able to identify the precise pitch features that people so readily and easily use. We have consequently developed a new technique for quantifying F0 features [2], in light of the basic psychophysics of pitch perception. In the present study, we have examined emotional speech in terms of the dissonance and harmonic tension among the pitches in the spoken sentences.

In preliminary experiments [2], we found that the distribution of F0 during normal speech of even short duration is frequently bimodal or trimodal (Fig. 1). Since descriptive statistics obtained from distributions that are not unimodal Gaussian are not informative, we have developed techniques to examine the multimodal substructure of the F0 spectrum without entering into the difficulties of the time domain (pitch contours).

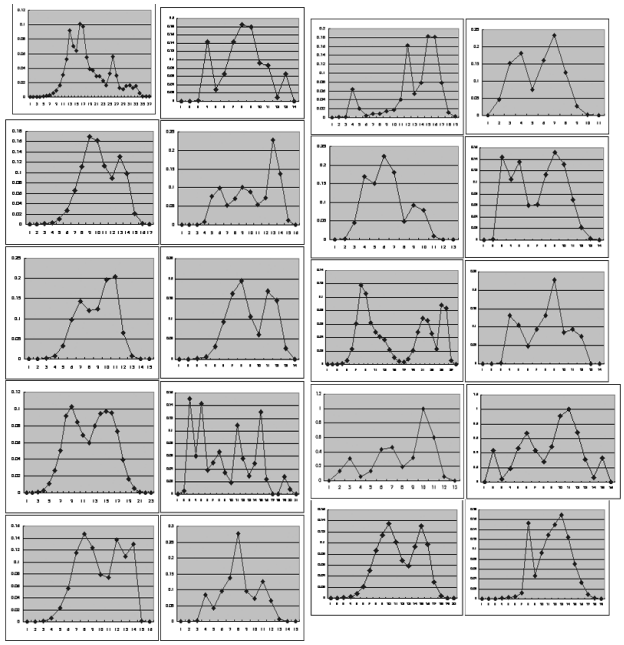

Figure 1: Typical examples of the multimodal pitch structure of 1-2 second samples of normal speech [2]. The vertical axis is intensity and the horizontal axis indicates pitch (in semitones). Details of the pitch substructure are unimportant in the present context, other than the fact that a description of "mean pitch" is meaningless. Clearly, each of these samples contains 2, 3 or more overlapping pitch components.

Our approach has been to abstract the fundamental frequencies of emotional speech during sentence-length utterances in order to examine the underlying interval/harmonic structure. This has been motivated from findings in the psychology of music perception that indicate that virtually all normal subjects can distinguish between consonant and dissonant intervals [3], between major and minor chords and between

resolved and unresolved chords [4]. We have therefore hypothesized [2] that the perception of the emotionality of normal speech relies upon this same inherent sensitivity to pitch combinations that human listeners exhibit with regard to music.

2. METHODS

In Experiment 1, we recorded "emotional" sentences (using the Praat software [5], 44100 Hz sampling rate) read by undergraduate subjects for partial credit in an introductory psychology course. The sentences described typical emotional events, such as a grandparent dying or finding money on the street, and were read "with empathy". The spoken sentences were therefore not "spontaneous speech", but the algorithm for calculating the affect of the speech can be applied to both music and spontaneous speech. For evaluation of the affect of the speech, each utterance was converted into unintelligible humming sequences, and then scored (in Experiment 2) for their positive or negative affect.

Pitch F0 was calculated at 1 millisecond intervals, giving 500-1000 pitch values per utterance. Those data were then used as input to an unsupervised "cluster" algorithm [6] that calculates a best fit between the raw data and the summation of 1-12 Gaussian clusters (radial basis functions). The number of clusters per utterance is determined automatically by a maximum entropy technique [6]. Each cluster has variable position and width along the frequency axis, and variable intensity (height). Figure 2 shows typical examples of the raw power spectrum and the best-fit sum of Gaussians.

Having detected several clusters with known frequencies and intensities, a dissonance model (Fig. 3A) similar to that of Plomp and Levelt [3] is used to calculate the total dissonance implied by the cluster structure, by summing the dissonance (D) of all cluster pairs (Eq. [1]).

$$D = \mu_A * c * (\exp(-a * x) - \exp(-b * x)) \qquad [1]$$

where μ_A is the mean amplitude of each pair of tones, a, b, and c are constants (1.20, 4.00 and 3.53, respectively) and x is the interval size (in semitones).

As is known from studies in music perception (7), the harmonic "instability" or "tension" of even simple 3-tone chords cannot be explained solely in terms of interval dissonance. To explain the resolved/unresolved character of chords, and the relative stability of major and minor chords [4], the total tension (T) of 3-tone combinations must also be considered. This can be done using the model shown in Fig. 3B and Eq. [2].

$$T = \mu_A * 2 * \mu_x * \exp(-(|x1 - x2| / d)^2) \qquad [2]$$

where μ_A is the mean amplitude of the 3 tones, μ_x is mean interval size (in semitones), x1 and x2 are the interval sizes (in semitones), and d is a constant (0.60).

The total "instability" (I) of the pitch curve can then be calculated as shown in Eq. [3]:

$$I = D + 0.1 * T \qquad [3]$$

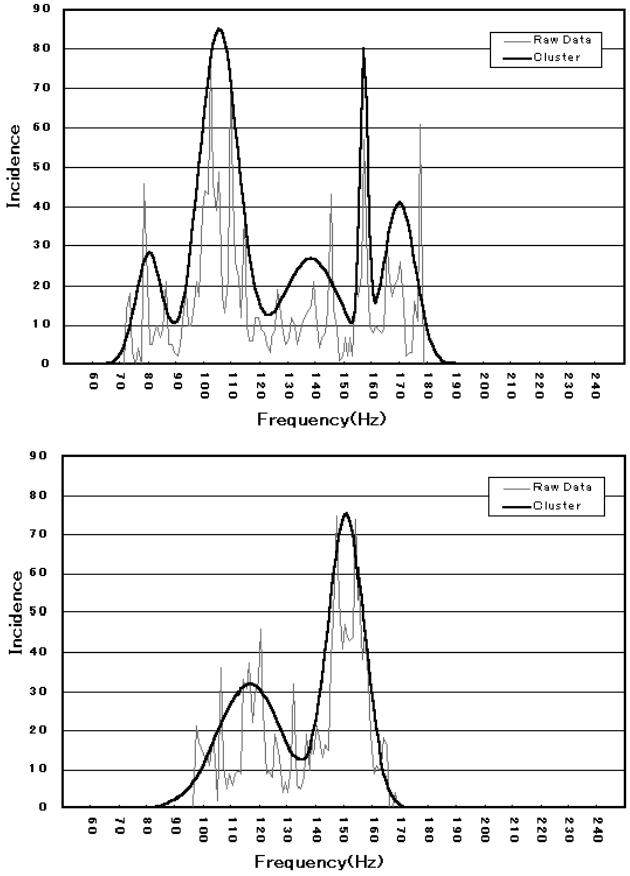

Figure 2: The pitch spectrum of utterances, together with the best fit obtained using several Gaussian clusters. Above is shown an example in which 5 distinct clusters were detected by the cluster algorithm; below is an example showing 2 pitch clusters.

In a musical context, Equation [3] has previously been shown to reproduce the experimental sequence of the evaluation of 3-tone chords [2]. Noteworthy is the fact that the dissonance and tension curves (Eqs. [1] and [2]) can be applied to pitch combinations that are unrelated to specific musical scales and unrelated to specific tuning systems (i.e., equitempered or just tuning). For this reason, the tonal "instability" calculation can be applied equally to the non-scalar pitches in normal speech and to the scalar tones in specific musical traditions.

As is seen in Figure 3, the calculation of the dissonance of pairs of pitches results in a relatively high value when they are approximately 1-2 semitones apart, and in a relatively low value when more than 2 semitones apart. The calculation of harmonic tension depends on the relative size of the two intervals in a 3-tone "chord". When the two intervals are approximately the same size, the calculated "tension" value is large, whereas two unequal intervals results in a low tension value. Musically, a state of high tension is perceived for augmented chords (two intervals of 4 semitones), whereas a major or minor chord (e.g., one interval of 3 semitones and one interval of 4 semitones) results in a stable, "resolved" or low tension value.

3. RESULTS

The total dissonance, tension and instability of the 144 speech utterances in Exp. 1 were calculated, and are shown in Table 1. Using the results of the evaluation of these utterances in Exp. 2, a second analysis was done on the 25% of utterances showing relatively strong "positive" affect (n=18) and the 25% showing relatively strong "negative" affect (n=18). These results are shown in Table 2.

Table 1: Tonal Analysis of the 144 Positive and Negative Utterances (1-3 second duration)

	Dissonance	Tension	Instability
Positive	0.440	1.527	0.593
Negative	0.567	2.002	0.767
t-value	-1.982	-1.744	-2.098
P-value	0.049	0.083	0.038

Table 2: Tonal Analysis of the 18 Most Positive and 18 Most Negative Utterances

	Dissonance	Tension	Instability
Positive	0.386	1.687	0.555
Negative	0.693	1.894	0.883
t-value	-2.110	-0.371	-1.745
P-value	0.042	0.713	0.090

It is seen that in both analyses there is greater dissonance among the clusters in the negative affect sentences, and consequently a greater "instability" of the pitch combinations.

Although we had anticipated a stronger effect for the utterances that had been selected for their relatively clear positive or negative affect, the results of the analysis showed greater statistical strength of the entire data set. In both cases the sentences with negative affect showed greater "instability", as would be predicted from the greater instability of minor chords, relative to major chords [2, 4, 7].

4. DISCUSSION

The ability of musically-untrained listeners to distinguish between major and minor chords, and between resolved and unresolved chords is a well-established, but truly amazing finding in the psychology of music perception. Similarly, the ability of normal listeners to detect the positive or negative affective state of a speaker, even when the meaning of the speech is unintelligible, is well-known from cross-cultural intonation studies – and indicates a sensitivity of the human ear to the information contained in, principally, the fundamental frequency of the voice. In light of the fact that isolated pitches and isolated intervals in music do not suffice to indicate the harmonic mode, we have attempted to deduce the affective "valence" of normal speech using a psychophysical model of harmony perception, i.e., using a model that does not rely on the concepts of traditional music theory, but can nonetheless be applied to musical phenomena. The present findings indicate that study of the interval and chordal substructure of normal speech may be a fruitful means for determining the positive or negative valence of "emotional" speech.

Our study of the pitch substructure of speech suggests that the missing component in current intonation theory is the consideration of relevant musical phenomena – specifically, harmony theory. That is, certain combinations of three (or more) tones carry with them the "universal" meanings of the major and minor modes. Whether the pitches are played simultaneously or sequentially, in music or in speech, the affective valence of certain combinations is apparent to most normal listeners. Therefore, the emotional "ring" apparent in both musical melodies and in speech prosody may have its origins in the same harmonic phenomena.

Although the psychophysical model of harmony perception outlined in Figure 3 is rather simple in comparison with the complexities of traditional harmony theory, it suffices to explain the basic pattern of perceived "sonority" or "harmoniousness" of 3-tone and 4-tone chords that has been reported in the music perception literature and that is, indeed, musical "common sense". That is to say, most normal listeners (both musicians and non-musicians) report that major chords are somewhat more harmonious than minor chords, and both major and minor chords are notably more harmonious than diminished and augmented chords [e.g., 2, 4]. This highly consistent finding in music psychology cannot be explained solely on the basis of the interval substructure of chords, even when upper partials are included, e.g., [7], but it is readily explained once the "tension" of certain 3-tone patterns are brought into consideration [2].

5. REFERENCES

[1] K.R. Scherer (1986). "Vocal affect expression." *Psychological Bulletin* 99, 143-165.

[2] N.D. Cook (2002) *Tone of Voice and Mind*, John Benjamins, Amsterdam.

[3] R. Plomp & W.J.M. Levelt (1965)." Total consonance and critical bandwidth." *Journal of the Acoustical Society of America 38*, 548-560.

[4] L.A. Roberts (1986). "Consonant judgments of musical chords by musicians and untrained listeners." *Acustica* 62, 163-171.

[5] P. Boersma & D. Weenink (2002). *Praat: A system for doing phonetics by computer*. (For details, see http://www.praat.org).

[6] C.A. Bouman (2002). "Cluster: An unsupervised algorithm for modeling Gaussian mixtures." (For details, see http://www.ece.purdue.edu/~bouman)

[7] R. Parncutt (1989) *Harmony: A psycho- acoustical approach,* Springer, New York.

[8] L.B. Meyer (1956) *Emotion and Meaning in Music*, University of Chicago Press, Chicago.

Acknowledgment: This work was supported by the "Research for the Future program", administered by the Japan Society for the Promotion of Science (Project No. JSPS-RFTF99P01401).

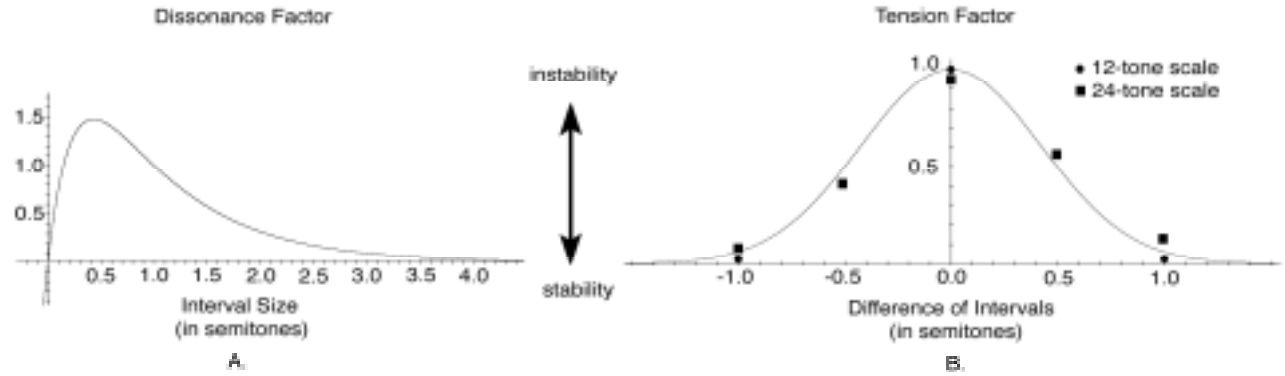

Figure 3: The dissonance curve (A) and the tension curve (B), the summation of which gives a total "instability" of multi-pitch combinations [2]. The dissonance curve is similar to those proposed by Plomp and Levelt [3] and others in the literature on pitch perception. The tension curve is explicitly a "3-body" effect and has not yet been incorporated into most accounts of harmony perception, but it is an important part of several qualitative discussions of modern music, notably, the work of Leonard Meyer [8]. Meyer noted that, whether sounded sequentially (as melody) or simultaneously (as harmony), the perception of intervals of equivalent size ("intervallic equidistance") is the source of the tension of the diminished and augmented chords and of the chromatic scales. By modeling the "tension", as in Equation [2], the affective quality of any number of pitch combinations – in music or in speech – can be quantified.