# Evaluation of the Affect of Speech Intonation Using a Model of the Perception of Interval Dissonance and Harmonic Tension

*Norman D. Cook, Takeshi Fujisawa and Kazuaki Takami*

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

## Abstract

We report the application of a psychophysical model of pitch perception to the analysis of speech intonation. The model was designed to reproduce the empirical findings on the perception of musical phenomena (the dissonance/consonance of intervals and the tension/sonority of chords), but does not depend on specific musical scales or tuning systems. Application to intonation allows us to calculate the total dissonance and tension among the pitches in the speech utterance. In an experiment using the 144 utterances of 18 male and female subjects, we found greater dissonance and harmonic tension in sentences with negative affect, in comparison with sentences with positive affect.

## 1. Introduction

It is a familiar idea that the intonation of speech is analogous to the melody of music, but all attempts to quantify pitch contours for "musical analysis" run into two formidable problems. The first is the identification of the "dominant" or "relevant" pitches from the continuous contours of fundamental frequency (F0) in normal speech. Several possible techniques are available (selecting the pitches that occur at peaks of intensity, averaging the pitch values during each vowel, etc), but, even if the dominant pitches can be determined, the tones of speech rarely fall at the frequencies of the notes of musical scales, making musical analysis all but impossible. As a consequence, it is essential first to develop a model of pitch perception which can handle tones that do not necessarily coincide with the tones of an arbitrarily-defined musical scale.

We have developed such a model – one that allows for a strictly psychophysical analysis of the basic phenomena of musical harmony – and have applied it to speech intonation. The model (Section 2) has two components, the first dealing with the consonance/dissonance of pitch intervals and the second dealing with the harmony/tension of three-tone pitch combinations. The model is applicable to various musical scales, but is designed explicitly to reproduce the empirically known evaluations of two- and three-tone musical phenomena in diatonic music – specifically, the dissonance of small (1-2 semitone) intervals, and the "tension" or "instability" of three-tone chords containing two equivalent intervals (e.g., diminished and augmented chords). We have previously reported on the musical applications of the model [1-7], but report here our first application to speech phenomena.

In order to identify the dominant pitches in short (1-3 second) utterances, we calculate the intensity-weighted frequency spectrum of the F0 for the entire utterance (the thin lines in Fig. 1), and then determine the location of the principal pitch components using an unsupervised "cluster" algorithm [8]. In effect, the algorithm fits radial basis functions (the thick Gaussian curves in Fig. 1) to any continuous data set. In the case of short utterances, 1-6 pitch clusters are normally found (most frequently 2 or 3 clusters). Advantages of the cluster algorithm include the fact that all of the pitch data from an utterance is used in determining the dominant pitches, and data-editing on the basis of linguistic theory is not required.

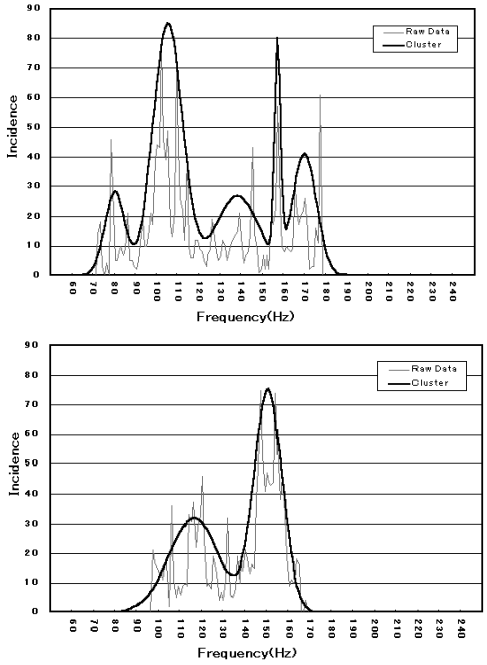


*Figure 1:* The pitch spectrum of short utterances, together with the best fit obtained using several Gaussian clusters. Above is shown an example in which 5 distinct clusters were detected by the cluster algorithm; below is an example showing 2 pitch clusters.

Various other techniques for simplifying the pitch data in normal speech might be used, but, in preliminary experiments in which the pitches in normal speech were "normalized" to the tones of musical scales, we have been struck by the high incidence of bimodal and trimodal pitch distributions typically obtained from sentence-length utterances (Fig. 2). Normalization of the pitch data to arbitrarily-defined musical scales (or to the arbitrarily-defined tone "levels" of conventional intonation theory), however, demands an unacceptable manipulation of the raw

data, so we favor the cluster method advocated by Bouman [8]. The algorithm employs an unsupervised maximum entropy technique that determines the smallest number of radial basis functions that can reproduce the raw data set. Although the raw data are greatly simplified with the cluster algorithm, there is no need for fitting the clusters to pre-defined pitch levels (scales).

Given the pitch location and height (amplitude) of the clusters, it is then a straight-forward matter to use those values to calculate (i) the total dissonance among the clusters (using a dissonance model similar to that of Plomp and Levelt [9]) and (ii) the harmonic tension of all three-tone combinations (using a newly developed psychophysical model of harmony [1]).

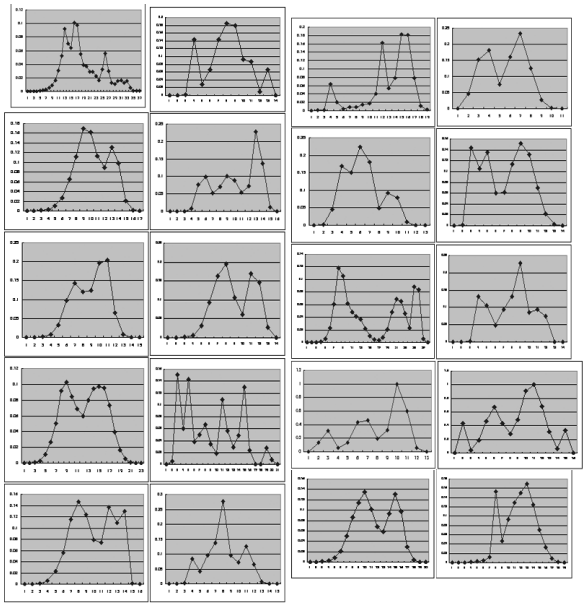

*Figure 2:* Typical examples of the pitch structure of normal 1-2 seconds speech utterances (from ref. [1]). The vertical axis is intensity and the horizontal axis indicates pitch (in semitones). Details of the pitch structure are unimportant in the present context, other than the fact that the mean and standard deviation of such multimodal distributions are not meaningful. Clearly, each of these samples contains 2, 3 or more overlapping pitch components.

## 2. Models for dissonance and tension

A pitch perception model (Fig. 3A) similar to those of Plomp and Levelt [9] and Sethares [10] was used to calculate the total dissonance between the tones detected in each spoken utterance. Given the frequency and amplitude of any number of tones, the dissonance (D) of all tone pairs can be computed as:

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

where $\mu_A$ 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 [11], the perceived sonority of even simple three-tone chords (the triads of harmony theory) cannot be explained solely in terms of interval consonance (and their upper partials). To explain the resolved/unresolved character of chords, and the relative stability of major and minor chords [12], the total tension (T) of three-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 $\mu_A$ is the mean amplitude of the three tones, $\mu_x$ is the mean interval size (in semitones), x1 and x2 are the interval sizes (in semitones), and d is a constant (0.60). As seen in Figure 3B, the model produces a high tension value when any three tones are spaced such that there are two intervals of approximately equal size. Musical examples of such "high tension" chords are the augmented and diminished chords, but Eq. [2] can also be applied to combinations of tones that do not fall on any scalar notes.

Given the total dissonance and the total tension of pitch combinations, the total "instability" (I) can then be calculated as shown in Eq. [3]:

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

where the constant (0.1) is used to adjust the relative contributions of interval dissonance and harmonic tension to the perceived instability.

The motivation for using psychophysical models for interval and harmony perception in the analysis of speech intonation is the fact that most normal subjects without musical training can distinguish between consonant and dissonant intervals, between major and minor chords and between resolved and unresolved chords [1, 9, 11, 12]. We have therefore hypothesized that the perception of the affect of normal speech relies upon this inherent sensitivity to pitch combinations that human listeners exhibit with regard to music. The principal difference between the pitch phenomena of musical melody and speech intonation is that melodies rely on conventional scales, whereas the pitches used in speech occur over a continuous range. For this reason, the preliminary theoretical task has been to devise psychophysical models that are independent of the scales of specific musical traditions, but that can nonetheless be applied to musical phenomena that employ scales.

## 3. Experiment

In an experiment employing 18 male and female Japanese undergraduates, we recorded 8 "emotional" sentences per subject (using the Praat software [13], 44100 Hz sampling rate). The sentences described typical emotional events, such as a grandparent dying or finding money on the street, and were read "with empathy". For evaluation of the affect of the speech, each utterance was converted into unintelligible humming sequences, and then scored for

their positive or negative affect by other undergraduates in a later session.

Pitch F0 was calculated at 1 millisecond intervals, giving 500-1000 pitch values per utterance. Those data were then used as input to the cluster algorithm [8] 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 [14], such that the minimum number of clusters is used. Each cluster has variable position and width along the frequency axis, and variable intensity (height). Figure 1 shows typical examples of the raw power spectrum and the best-fit sum of Gaussians.

In a musical context, Eqs. [1-3] have previously been shown to reproduce the experimental sequence of the evaluation of three-tone chords [1]. 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 (e.g., 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 various 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 0.5-1.5 semitones apart, and in a relatively low value when more than 2.0 semitones apart. In contrast, the calculation of harmonic tension depends on the relative size of the two intervals in a three-tone triad. When the two intervals are approximately the same size, the calculated "tension" value is large, whereas two unequal intervals result in a low tension value. Musically, a state of high tension is perceived for, e.g., augmented chords (two intervals of 4 semitones), whereas major or minor chords (e.g., one interval of 3 semitones and one interval of 4 semitones) result in a stable, "resolved" chord with low tension value.

## 4. Results

The total dissonance, tension and instability of the speech utterances with positive or negative affect were calculated, and are shown in Table 1. It is seen that there is greater dissonance and greater tension among the clusters in the negative affect sentences, and consequently a greater total "instability" of the pitch combinations, in comparison with the sentences read with positive affect.

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

| | Dissonance | Tension | Instability |
|---|---|---|---|
| | Eq. 1 | Eq. 2 | Eq. 3 |
| 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 |

## 5. Conclusions

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 intonation studies [15] – 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 neither isolated pitches nor isolated intervals in music suffice to indicate the harmonic mode (specifically, the major or minor mode), we have hypothesized that the perception of affect in speech requires consideration of three-tone combinations. We have therefore attempted to deduce the affective "valence" of normal speech using a psychophysical model of harmony perception, i.e., 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 examination of the three-tone (chordal) substructure of normal speech may be a fruitful means for determining the positive or negative valence of "emotional" speech. Stated negatively, the importance of harmonic structure in determining the affective mode of music is strong indication that first-order statistics on the F0 of speech cannot explain the affective quality of voice intonation and, moreover, that intonational systems that do not go beyond interval effects (rising or falling pitches) cannot in principle capture the affect of voice. Similarly, although mean pitch, standard deviation of pitch and all kinds of information on rising and falling pitch intervals cannot explain musical mode, three-tone combinations provide unambiguous information.

Although the psychophysical model of harmony perception outlined in Fig. 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 three-tone chords that has been reported in the music perception literature and that is, indeed, musical "common sense" (in the Western diatonic tradition). 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., 1, 12]. 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., [11], but it is readily explained once the "tension" of certain three-tone patterns are brought into consideration [1].

Our study of the pitch structure of speech suggests that the missing component in current intonation theory is the consideration of the relevant three-tone musical phenomena, i.e., harmony. (Indeed, the psychophysics of three-tone chords is missing from many models of musical pitch perception [e.g., ref. 16], but without consideration of

three-tone effects, it is known that the relative sonority of the triads of traditional harmony theory cannot be explained on the basis of intervals and upper partials [e.g., ref. 11]). From studies in music perception, it is known that certain combinations of three (or more) tones carry with them the "universal" (or at least extremely widespread) meanings of the major and minor modes. As a consequence, whether pitches are played simultaneously or sequentially, in music or in speech, the affective valence of certain combinations is apparent to most normal listeners and may therefore have its origins in the same harmonic phenomena.

## 6. Acknowledgement

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).

## 7. References

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

[2] Cook, N.D., "Explaining harmony: The roles of interval dissonance and chordal tension." *The Biological Foundations of Music.* New York: *New York Acad. Sci.*, Vol. 930, pp. 382-385, 2001.

[3] Cook, N.D., "Tonal harmoniousness is determined by two distinct factors: interval dissonance and chordal tension." *Proc. 5th Int. Conf. Music Percept. Cogn.*, Keele, 2000.

[4] Cook, N.D., Callan, D.E., & Callan, A., "An fMRI study of resolved and unresolved chords," *Proc. 6th Ann. Meet. Soc. Music Percept. Cogn.*, Kingston, 2001.

[5] Cook, N.D., Callan, D.E., & Callan, A., "Frontal areas involved in the perception of harmony." *Proc. 8th Int. Conf. Funct. Mapping Human Brain*, Sendai, 2002.

[6] Cook, N.D., "The psychoacoustics of harmony: Tension is to chords as dissonance is to intervals" *Proc. 7th Int. Conf. Music Percept.Cogn.*, Sydney, 2002.

[7] Cook, N.D., Callan, D.E., & Callan, A., "Frontal lobe activation during the perception of unresolved chords." *The Neurosciences and Music*, Venice, 2002.

[8] Bouman, C.A., "The unsupervised cluster algorithm" http://www.ece.purdue.edu/~bouman, 2002.

[9] Plomp, R. & Levelt, W.J.M., "Total consonance and critical bandwidth." *J. Acoust. Soc. Amer. 38*, 548-560, 1965.

[10] Sethares, W. A., *Tuning, Timbre, Spectrum, Scale*. Springer, New York, 1999.

[11] Parncutt, R., *Harmony: A psychoacoustical approach*. New York: Springer, 1989, pp. 140-142.

[12] Roberts, L.A., "Consonant judgments of musical chords". *Acustica* 62, 163-171, 1986.

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

[14] Akaike, H., "A new look at the statistical model identification," *IEEE Trans. Automat. Control* AC-19, 716-723, 1974.

[15] Scherer, K.R., "Vocal affect expression." *Psych. Bull.* 99, 143-165, 1986.

[16] Lerdahl, F., *Tonal Pitch Space*, Oxford University Press, New York, 2001, pp. 80-81, 142,.

[17] Meyer, L.B., *Emotion and Meaning in Music*. University of Chicago Press, Chicago, 1956, pp. 163-167.

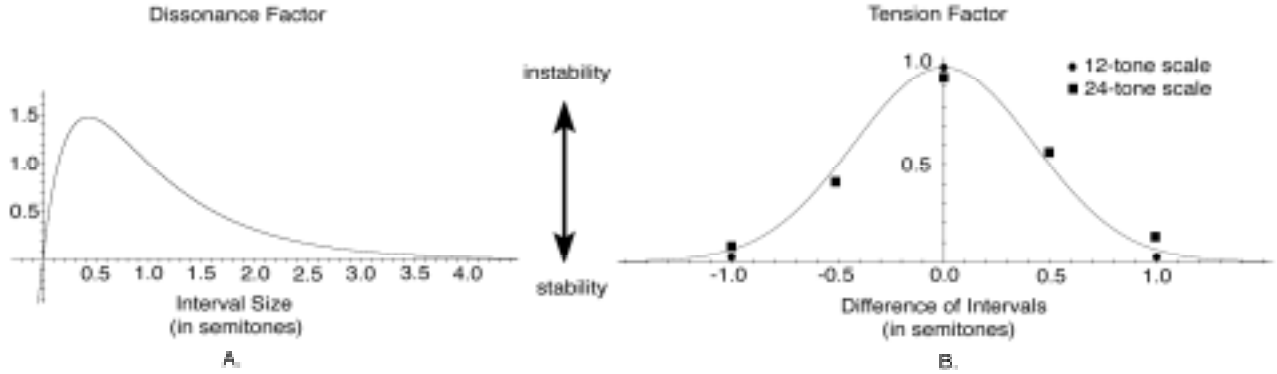


*Figure 3:* The dissonance curve (A) and the tension curve (B), which can be used to calculate the total "instability" of multi-pitch combinations [1]. The dissonance curve is similar to those proposed by Plomp and Levelt [9], Sethares [10] and others in the literature on pitch perception. The tension curve is explicitly a "three-body" effect. The data points are from experiments using 12-tone and 24-tone scales, discussed in ref. [1]. The crucial role of three-tone effects (i.e., chord perception) is an important part of many qualitative discussions of music perception, most notably, the work of Leonard Meyer [17]. Meyer argued 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 harmonic "tension", as in curve B (Eq. [2]), the affective quality of any number of pitch combinations – in music or in speech – can be quantified.