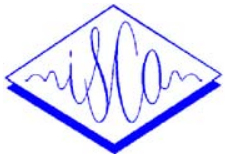

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Application of a Psychoacoustical Model of Harmony to Speech Prosody

Norman D. Cook, Takashi Fujisawa and Kazuaki Takami

Department of Informatics
Kansai University, Osaka, Japan
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 4 quasi-musical measures: dissonance, tension, total harmonic "instability", and "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 harmony measure is essential for determining the affective character of the tone of voice.

1. Introduction

The central paradox in the study of emotional prosody is the fact that affective information is carried in the fundamental frequency (F0) of the voice, and 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 [1]. 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?

We have approached this question by returning to the basic psychophysics of music perception in order to determine what kinds of pitch patterns that people (both musicians and non-musicians) normally perceive in music. We have been able to develop a psychoacoustical model of musical harmony that solves two ancient problems in music theory [2-10]. The first is why there exist both resolved and unresolved chords – i.e., why some tonal combinations are perceived as stable, harmonious and final, while other tonal combinations are perceived as tense, inharmonious and incomplete. The second problem is why, among the harmonious chords, there are some that have the dark, negative, "sad" ring typical of the minor modality and others that have the bright, positive, "happy" ring of the major modality. By consideration of the psychoacoustics of both 2-tone dissonance and 3-tone tension, our model allows for quantitative measures of harmonic modality without relying on the concepts of traditional harmony theory. In other words, the model can be applied both to scalar pitch systems (e.g., the diatonic scales of Western music) and to non-scalar systems (e.g., the continuous pitch changes in typical speech utterances). The chief merit of the model is that it does ***not*** rely on the alignment of pitches to any particular musical scale (diatonic or otherwise), and can therefore be applied directly to the pitch events in speech [10-13]. Here we outline the musical model and report the results of its use in a speech prosody experiment.

Details of the model are available in several recent publications, but the underlying ideas and the application to speech prosody will be presented here. The most important insight of the model is borrowed from music perception. That is, it is known that at least ***three*** tones must be heard to perceive the major or minor modality of music (whether heard sequentially as a melody or simultaneously as a harmony). This fact implies that the study of voice intonation cannot rely solely on ***two***-tone interval effects (e.g., rising pitch or falling pitch), but must consider the relative pitch heights of at least three-tone combinations.

2. Model

We have found that both problems concerning resolved and unresolved harmonies and concerning major and minor harmonies can be resolved 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 [14-16]. 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" [17]. 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]. The total instability, I, of the tone combination can then be calculated as the sum of these two factors [Eq. 3].

$$D = minAmp * c * (\exp(-a * x) - \exp(-b * x)) \quad (1)$$

$$T = minAmp * \exp(-((x1 - x2) / d)^2) \quad (2)$$

$$I = D + e * T \quad (3)$$

10.21437/SpeechProsody.2004-35

where *minAmp* indicates the amplitude of the pitch with the smallest amplitude, *x*, *x1* and *x2* are interval sizes (in semitones) and *a~e* are constants (1.20, 4.00, 3.53, 0.60, 0.10, respectively) chosen to produce the known (experimentally measured) relative sonority of the triads (major > minor > diminished > augmented). The 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).

By calculating both the dissonance and the tension of tone combinations, it is found that the empirically-known sequence of tonal "stability" ("harmoniousness", "sonority", etc.) is easily reproduced, whereas interval-based (dissonance only) models inevitably have difficulties in explaining the perceived instability of the augmented chord (Table 1).

TABLE 1
THEORETICAL TENSION AND MODALITY SCORES FOR THE TRIADS OF TRADITIONAL DIATONIC MUSIC
(Calculations were made using the first three partials with amplitude decreasing as 1/n, but the results are not highly sensitive to the number or strength of the upper partials [2, 7].)

CHORD	Tension Score	Modality Score
Major		
4-3 Root	0.990	2.532
3-5 1st Inversion	0.940	1.096
5-4 2nd Inversion	1.351	3.198
Minor		
3-4 Root	1.021	-2.588
4-5 1st Inversion	1.320	-1.447
5-3 2nd Inversion	0.940	-1.447
Diminished		
3-3 Root	3.019	-0.748
3-6 1st Inversion	2.693	0.066
6-3 2nd Inversion	2.693	-1.120
Augmented		
4-4 Root	4.564	1.261

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. 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, *x1* and *x2*, in a three-tone chord [Eq. 4].

$$M = f * minAmp * (x1 - x2) * \exp(-(x2 - x1)^2 / 2) \qquad (4)$$

where *f* is a constant (1.65), the intervals, *x1* and *x2*, are defined in semitones; *x1* is the lower interval and *x2* is the higher interval. The modality curve is shown in Figure 1C.

The meaning of the three curves 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 ~±1.0) have low tension values. (C) When the lower of the two intervals in a triad is larger than the upper interval, the modality is score is positive (major-like); when the lower interval is smaller, the modality score is negative (minor-like). The simple curves in Figure 1 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 Table 1).

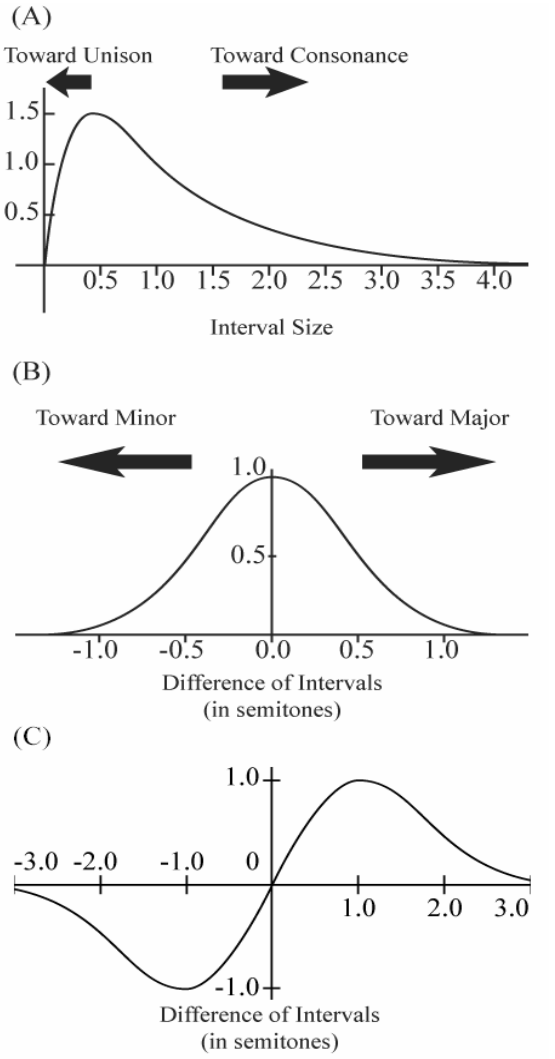

Figure 1: *The three main factors contributing to the perception of harmony. (A) shows the dissonance factor, computed for every pair of tones and their upper partials. (B) shows the tension factor, computed for every tone triplet. (C) shows the modality factor that gives a positive (major chord-like) 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). See refs. [2-13] for further details.*

3. An Intonation Study

The psychoacoustic model outlined above was designed specifically to reproduce the experimental results concerning the perception of diatonic chords [18]. Its significance for harmony theory has been discussed elsewhere [2-10], but we have recently applied these same formula to the pitch phenomena of emotional speech. In the first experiment [10-13], 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 [19]), 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 regardless of the speaker's original intention.

TABLE 2
THE CONFUSION MATRIX
(Note that a majority of the utterances with intended positive or negative affect were correctly identified as such. The total number of utterances is 204, rather than 216, because 12 utterances did not show multiple pitch clusters, and were therefore inappropriate for harmonic analysis.)

		Perceived Affect			Total
		Pos33%	Amb33%	Neg33%	
Intend-ed Affect	Positive	40(59.7%)	23(34.3%)	4(6.0%)	67
	Tension	16(22.9%)	25(35.7%)	29(41.4%)	70
	Negative	12(17.9%)	20(29.9%)	35(52.2%)	67

A confusion matrix was calculated using the speaker's intended affect (positive, negative or ambivalent) and the perceived affect in the utterance evaluation experiment. As shown in Table 2, 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. 2, 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 [20] – normally 3-5 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)-(4) (details of the technique are provided in ref [11]).

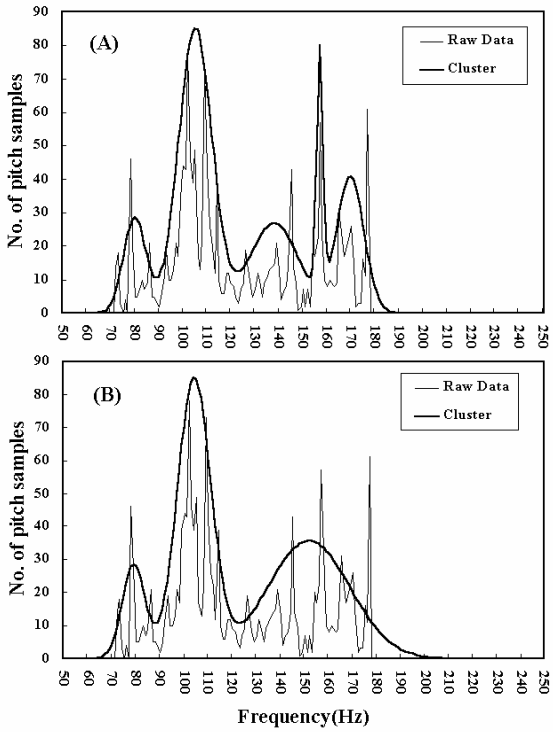

Figure 2: *A pitch histogram and its reconstruction using the cluster algorithm [19]. The fitting of Gaussian clusters (thick lines) to the raw data (thin lines) is attempted with 1-12 Gaussian curves and the "best fit" (A rather than B, in this case) is selected on the basis of a maximum entropy calculation [20]. The modality (etc.) of the pitches in the utterance was then calculated using the 5 pitches indicated by the 5 peaks (at 81, 108, 139, 158 and 170 Hz) in (A).*

The mean dissonance, tension, instability and modality of the 204 speech utterances of the two types of perceived affect were calculated. There was greater dissonance among the clusters in the Negative Affect sentences (df=190, t=-2.464, $p<0.0146$) and consequently a greater "instability" of the pitch combinations. These effects reflect the fact that negative affect is typically expressed with a smaller range of F0 [19], and therefore necessarily a higher dissonance value. Significant differences in tension between the 2 conditions were not found. The most interesting results concern modality. The utterances perceived as having positive affect showed higher modality values, indicative of greater major-like pitch substructure (df=185, t=2.01, $p<0.046$). The prediction that the sentences with negative affect would have negative (minor-like) modality scores was not found, but the relatively lower values indicates the anticipated, less positive pitch structure.

TABLE 3
THE RESULTS OF FACTOR ANALYSIS

	Factor 1	Factor 2	Factor 3	Factor 4
F0 Range	**.916**	.148	.168	-.125
F0 St. Dev.	**.913**	.235	.233	.077
F0 Max.	**.914**	.253	.128	.357
F0 Min.	.031	.178	-.068	**.951**
F0 mean	**.675**	**.499**	.214	**.613**
Duration	-.364	**-.918**	-.193	-.243
Rate	.108	**.905**	.054	.218
Tension	.389	.023	**.788**	-.097
Dissonance	-.145	.098	**.763**	.013
Modality	.284	.158	**.416**	.051
Correlation	Factor 1	0.271	0.207	0.166
	Factor 2		0.137	0.308
	Factor 3			-0.006

A second result of the intonation experiment is shown in Table 3. That is, factor analysis using 10 acoustical features revealed 4 independent factors contributing to the evaluation of the utterances. Importantly, the quasi-musical features of the present model were found to form a factor distinct from the conventional acoustical measures of the voice. In other words, our measures of tension and modality were not simply another way of calculating mean frequency, range, etc., but rather constitute an independent measure of the relative use of multiple pitch combinations, i.e., the "harmony" of speech. The prosody experiment itself was only a partial success, because the correlation between perceived affect and the modality scores was statistically significant ($p<0.05$), but low ($R=0.212$). Reasons for the limited success of the experiment are numerous and will be addressed in future work.

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 [22]. 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.

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). Aspects of this research have previously been published in refs. [2-13].

7. References

[1] Scherer, K.R. (1995) Expression of emotion in voice and music. *Journal of Voice* 9, 235-248.

[2] Cook, N.D. (2000) Chordal harmoniousness is determined by two distinct factors: interval dissonance and chordal tension. *Proc. 6th Inter. Conf. Music Percep. Cogn.*, August, Keele.

[3] Cook, N.D. (2001) Understanding harmony: the role of chordal tension. *Ann. New York Acad. Sci.* 930, 382-385.

[4] Cook, N.D. (2002) An fMRI study of resolved and unresolved chords, *Proc. 6th Ann. Meet. Soc. Music Percept. Cogn.*, July, Kingston.

[5] Cook, N.D. (2002) The psychoacoustics of harmony: Tension is to chords as dissonance is to intervals. *Proc. 7th Inter. Conf. Music Percep. Cogn.*, July, Sydney.

[6] Cook, N.D., Callan, D.A., & Callan, A. (2002) Frontal areas involved in the perception of harmony. *8th Inter. Conf. Func. Mapping Human Brain*, June, Sendai, Japan.

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

[8] Cook, N.D., Fujisawa, T., & Takami, K. (2003) A functional MRI study of harmony perception, *Meet. Soc. Music Percep. Cogn.*, June, Las Vegas.

[9] Fujisawa, T., Takami, K., & Cook, N.D. (2003) On the role of pitch intervals in the perception of emotional speech. *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, April, Tokyo.

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

[11] Cook, N.D., Fujisawa, T., & Takami, K. (2004) Evaluation of the affective valence of speech using pitch substructure. *IEEE Speech & Signal Processing* (in press).

[12] Fujisawa, T., Takami, K., & Cook, N.D., 2003. On the role of pitch intervals in the perception of emotional speech. *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, April, Tokyo, pp.231-234.

[13] Cook, N.D., Fujisawa, T., & Takami, K. (2003) Evaluation of the affect of speech intonation using a model of the perception of interval dissonance and harmonic tension, *Eurospeech 2003*, Sept., Geneva.

[14] Plomp, R. & Levelt, W.J.M., 1965. Total consonance and critical bandwidth. *JASA* 38, 548-560.

[15] Kameoka, A., & Kuriyagawa, M., 1969. Consonance theory (Parts I and II) *JASA* 45, 1452-1469.

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

[17] Meyer, L., 1956. *Emotion and Meaning in Music,* Chicago University Press, Chicago.

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

[19] Boersma, P., & Weenink, D., 2003. *Praat: a system for doing phonetics.* www.praat.org

[20] Bouman, C.A., 2002. *Cluster: an unsupervised algorithm for modeling Gaussian mixtures.* www.ece.purdue.edu/ ~bouman

[21] Rissanen, J., 1983. A universal prior for integers and estimation by minimum description length. *Annals of Statistics* 11, 417-431.

[22] Wallin, N.L., Merker, B., & Brown, S. (eds.) (2000) *The Origins of Music*, MIT Press, Cambridge, Mass.