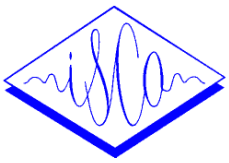

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

EMOTIONAL MODALITIES AND INTONATION IN SPOKEN LANGUAGE

Cari Spring and Donna Erickson

Speech and Hearing Science
Ohio State University
Columbus, Ohio 43210 USA

Thomas Call

TACAN Corporation
2330 Faraday Avenue
Carlsbad, California 92008 USA

ABSTRACT

This study reports on listener perception of contrastive emphasis in natural language utterances varied for emotional modality. Length and F0 correlates of well-perceived corrected digits indicate that length but not a L+H* rising contour cues listeners that the digit is emphasized. Length of vowels poorly perceived for contrastive emphasis but agreed by listeners to be emotional are compared with well-perceived corrected digits. Utterances which listeners agree are emotional are shown to have a high misperception rate for the corrected digit, and those with low listener agreement for emotion have a low misperception rate for the corrected digit.

10.21437/ICSLP.1992-227

1. INTRODUCTION

Pierrehumbert and Hirschberg [1] identify the L+H* pitch accent as the contour used by speakers to convey that the accented item is contrasted or corrected with some alternative. Westbury and Fujimura [2] show that the articulatory realization of contrastive emphasis is a large amplitude of jaw movement on the emphasized item. Erickson (pc) reports that the contrastive emphasis found in Westbury and Fujimura's corpus is also realized invariantly with the L+H* tonal contour. Previous studies of correction, contrast or hereafter, *contrastive emphasis,* are based on isolated spontaneous speech acts, for which the researcher intuitively recognizes that the intonational contour functions as contrastive emphasis, or upon read, laboratory, speech. Thus, while identifying the acoustic manifestations of contrastive emphasis as L+H* and length (which we assume follows from large mandible excursion), previous studies are based on data whose emotional modality (also *affect, speaker affect, emotional affect,* or *emotion*) is suppressed or absent in the utterance, or on spurious, natural language utterances. In a corpus of natural language data where contrastive emphasis occurs in utterances with varied emotional modality: i) if the verbal expression of emotion is added to the intonated morphosyntactic representation, here the contrastively emphasized expression, then the length and F0 signifying contrastive emphasis should remain constant under emotional modality variation; or ii) if emotion interacts with intonation and/or metrical structure of the morphosyntactic phrase, the acoustical correlates of contrastive emphasis should vary under conditions of varied emotional modality (cf. Scherer et al [3]).

2. CONTRASTIVE EMPHASIS IN NATURAL LANGUAGE

Four speakers provided the natural language corpus. Subjects were asked where s/he or another person lived or worked. A visual computer prompt provided the information. The subject was instructed to answer the questions in as natural a manner as possible (see Erickson and Fujimura [4] for further details of purpose and methodology of the study). In cases where the subject was asked where s/he lived, the elicitor would repeatedly misunderstand the subject's response by repeating back the digit sequence with the initial, medial, or final digit incorrect. Such a sequence is referred to here as a *dialogue*, and is exemplified in figure 1.

The interaction between the researcher and the subject is an *exchange*, and the response of the subject in a given exchange is an *utterance* (researcher's utterances are parenthesized). Thus there are four utterances in 1: the initial report given by the subject is a *first*

Exchange 1. (Where do you work?)
a. I work at 5 5 9 Pine Street.
H* H* !H* !H* !H* L- L%

Exchange 2. (That was 959 Pine Street?)
b. No, not 959. It's 5 5 9.
H* H-H% H*H- H* L- L%

Exchange 3.(I'm sorry, I'm not hearing...that was 959 P.S?)
c. No. 5 5 9.
H*H-H% H* !H* L- L%

Exchange 4. (I've got it...it was 959 Pine Street, right?)
d. No, that's wrong. It's 5 5 9.
H*H-H% H*L- H* L- L%

Figure 1: Transcription of Subject 1's DIALOGUE 10.

utterance, as in 1a; it is a *no-clarification* utterance. The *second utterance* occurs in the second exchange, where the subject gives the *first clarification*, 1b.The *second clarification* is the *third utterance*, 1c, and the *third clarification* is the *fourth utterance*, 1d. Dialogues where the initial digit is the focus of the misunderstandings are *initial-corrections* (only one digit is misunderstood per dialogue; eg. in figure 1 the initial digit is always misunderstood so 1b-d, the second, third, and fourth utterances, are *initial-corrections*), dialogues where the medial digit is misunderstood are *medial-corrections*, *final-corrections* are dialogues in which the final digit is misunderstood. This context of excessive misunderstanding was used to elicit the emotional modalities of natural language in a laboratory setting.

Utterances elicited from subject 1 (hereafter S1) are the focus of the remainder of this paper. While S1 uses the characteristic L+H* on the corrected item in some clarification utterances [4], it is not used consistently. Returning to figure 1b-d, the corrected initial digit is marked by a high pitch accent and high phrase and boundary tones, which resulted in a rising tone on each corrected digit in dialogue 10 (we thank Mary Beckman for transcription assistance). In other utterances, prosodic characteristics--including F0 range and vowel length--varied greatly.

To test the ability of listeners to identify items with contrastive emphasis in emotionally varied utterances and to identify characteristic properties of well-perceived utterances, 60 randomized utterances from the S1 corpus were played to listeners.

3. PERCEPTION OF CORRECTED DIGITS

Ten native speakers of English listened to 60 tape recorded randomized utterances of S1 on WAVES+ ESPS 3, four times. The first time, listeners were asked to identify whether there was correction on any digit, and if so, to identify the location of the correction. On the second pass listeners identified which emotion, if any, was present in the utterance. On the third pass, listeners were instructed to rate the intensity of emotion (not further discussed in this paper). The fourth time, listeners re-identified just the corrected items, i.e. repeated the first task, with no access to previous answers.

The results of two listeners were dismissed from this study:

one listener's second identification of contrastive emphasis was extremely poor and one was very good; both varied from the mean by more than two standard deviations. For the remaining eight listeners, there were 16 total listening judgements (2 listening passes x 8 listeners) for contrastive emphasis in 57 utterances (three utterances did not have a well-defined set of three digits and were excluded from the analysis). The total number of utterances missed, by error rate, is shown in figure 2.

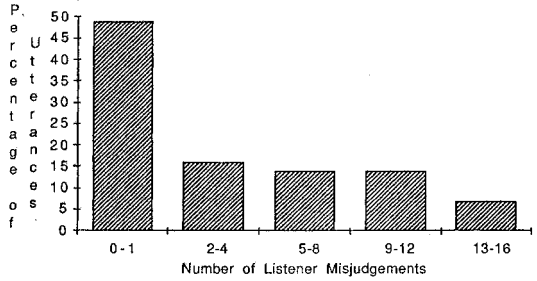

Figure 2. Error rate of misperceptions by number missed. Percentage shown is percent of 57 utterances which had the number of errors indicated (out of 16 listening judgments per utterance).

Figure 3 shows overall percent of corrected items missed, by location of errors: initial-corrections were missed most (44%), medial-corrections were second most missed (31%), no-correction utterances were seldom missed (11%) and final-corrections were missed least (7%).

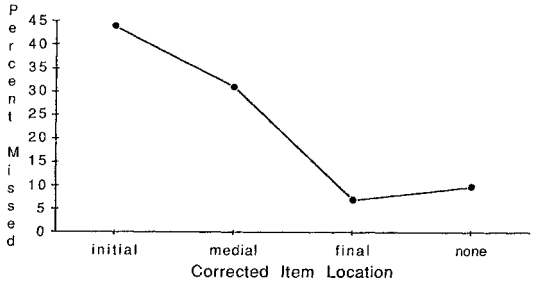

Figure 3: Percent of corrected items (*CI*) missed, by location of CI.

To test whether length and the L+H* tone are characteristic components of well-perceived corrected digits, length and F0 contours of well-perceived no-correction utterances, final-corrections (best perceived category) and initial-corrections (worst perceived) were examined (medial-correction data are not reported in this preprint).

The average vowel length of initial, medial, and final digits in well-perceived no-correction utterances (i.e. utterances perceived correctly 15-16/16 judgements) are reported in figure 4a. Length was measured using the wave form and spectrographs of the WAVES+ ESPS 3 software program. As shown, the final digit averaged 57/34 ms longer than the medial and final digits, respectively.

	initial	medial	final
a. $\bar{x}$	189	166	223
b. $\bar{x}$	130	127	250
c. $\bar{x}$	305	210	230

Figure 4. Average vowel length (in milliseconds) of initial, medial, and final digits in well-perceived, no-correction utterances, 4a, final-correction utterances, 4b, and initial-correction utterances, 4c.

Typically each digit in well-perceived, no-correction utterances has a H* pitch accent, and digits occur in one phrase, as exemplified by tracings in figure 5.

Final-correction utterances are the best perceived of all utterances, with 10 out of 13 final-correction utterances perceived correctly virtually 100% of the time. Average vowel length of the initial, medial and final digits in well-perceived final-correction

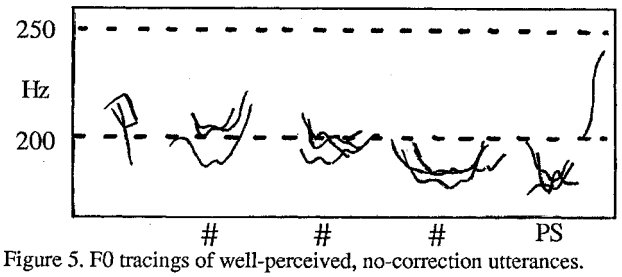

Figure 5. F0 tracings of well-perceived, no-correction utterances.

utterances is given in figure 4b. As shown, final-correction utterances are characterized with substantial length on the final digit, averaging ~120 ms longer than other digits. (Several initial, medial and no-correction utterances were perceived as mostly or entirely final-corrections; listener error toward a final-correction decision appears to be highly influenced by vowel length on the final digit.)

At the same time that the final digit of final-corrections is very long relative to the other digits, and as compared to the final digit in well-perceived, no-correction utterances (figure 4a), these final digits generally have a rapid falling tone and a directly adjoining phrase and boundary tone. They have the same general F0 range across the three digits as well-perceived no-correction utterances (figure 5), but the F0 ends lower on the final digit than on no-correction counterparts and intra-digit F0 is generally a falling tone, as shown in figure 6. These well-perceived final-corrections do not generally utilize the L+H* pitch accent (but note the one rising boundary tone in 6).

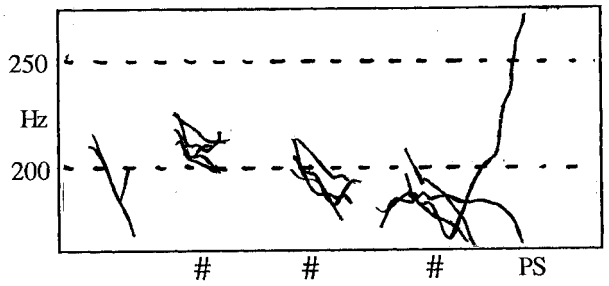

Figure 6. F0 tracings of well-perceived final-correction utterances.

Initial-corrections were most missed. No initial-corrections were identified as such by all listeners; rather the best perception rate for these types had a 3-4/16 misjudgement error rate. S1 uses the rising tone indicating contrastive emphasis [1] most often on initial-corrections; the rise is often extreme, and phrase and intonation boundaries generally follow the corrected digit. The F0 pattern is most often characterized as a high or rising pitch accent followed by upstep--i.e. a phonetic pattern arising from the tonal sequence of H* or L+H* followed by H- H% on the initial digit [1], [5]. F0 tracings of initial-correction utterances (including both 'fairly-perceived', 3-4/16 misperception rate, and 'poorly-perceived', 8-15/16 misperception rate) are shown in figure 7 (figure 1b-d shows a typical analysis of the rising tone on the initial digit).

Even though initial-corrections contain the corrected digit most often associated with a rising tone, they are relatively poorly perceived as initial-corrections. The invariant property of the two fairly well-perceived initial-corrections (3-4/16 misjudgments) was

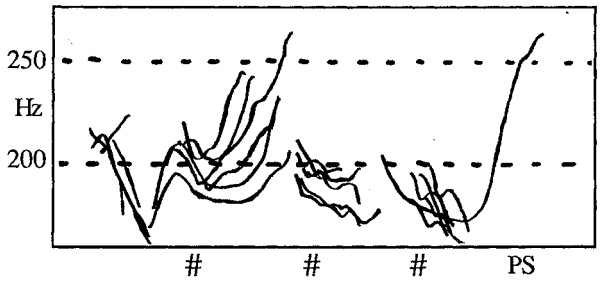

Figure 7. F0 tracings of initial-correction utterances.

vowel length on the initial digit relative to the other digits. Figure 4c shows the average length of the initial, medial and final digits on fairly well-perceived initial-correction utterances; the initial digit averages 95/75 ms longer than medial and final digits, respectively. That vowel length cues for contrastive emphasis is illustrated by comparing dialogue 6.2 with 16.2 in figure 8, where the former is fairly well perceived (3/16 error rate) and the latter is fairly poorly perceived (8/16 error rate). Utterance 6.2 has a slight rising tone, while 16.2 has an extreme rising tone on the initial digit, however it appears that relative length, not tone, is used to identify correction in the well-perceived utterance: in 6.2 the initial digit is 90 ms longer than the final digit; in 16.2 the initial digit is just 20 ms longer than the final digit.

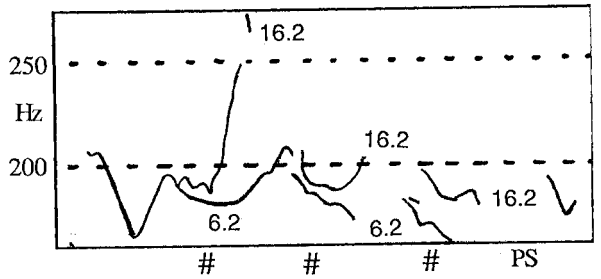

Figure 8. F0 tracings of utterance 6.2--well-perceived--and 16.2--poorly perceived initial-corrections.

4. EMOTIONAL MODALITY AND CONTRASTIVE EMPHASIS

To test listener agreement for presence of emotion in utterances, listener agreement on quality of emotion, and correlation between the presence of emotion and the ability to identify contrastive emphasis, the eight listeners judged the emotion on 57 randomized utterances. Vowel length was measured and F0 contours examined for a subset of utterances where emotion was agreed upon and contrastive emphasis was poorly perceived by listeners.

A judgement of *emotion absent* was assumed to be indicated by by three terms: *none, normal, calm*, while a judgement of emotion present was indicated by all other terms (figure 13 provides a list of all terms employed by listeners). Five total judgements were excluded from the analysis (each occurred once): *none-happy, none-irritated, none-annoyed, sad, unhappy*. Agreement on the presence of emotion for a given utterance was assumed to be indicated when one or fewer listeners disagreed with the remainder of the group about the presence or absence of emotion, and disagreement was taken to be indicated when two or more listeners disagreed with the others.

Listeners agreed on the presence/absence of emotion for 35 of the 57, and disagreed on 22 of the 57 utterances. Utterances with a 0-8/16 error rate on the identification of the corrected item disagree about the emotional content in 18/45 utterances, or 40% of the time, while utterances with an error rate of 9-15/16 times on the corrected item disagree on the presence of emotion 4/12 times, or 33% of the time. These figures show that utterances with the corrected item mis-identified tend to have the emotion better perceived, while utterances for which the corrected item is perceived well tend to have less agreement about the emotional content. In fact, most utterances having few or no misjudgements on the corrected item but with disagreement for emotion are no-clarification or first-clarification utterances, while second and third-clarification utterances are most agreed upon for emotion but most missed for contrastive emphasis, as shown in figure 9. No-clarification utterances are least likely to be missed for corrected item (listeners know that these are no-clarification items), but are most likely to have disagreement over the emotional content of the utterance. First-clarifications are next most likely to have the corrected digit correctly perceived, but are also next most likely to have poor agreement on the emotional content of the utterance, and so on.

These results are not surprising: speakers might be expected to become more emotional as they excessively reclarify linguistic information, and listeners agree that these utterances are most highly emotionally charged. Of the 8/12 utterances for which listeners agree on the emotional content of third and fourth utterances, they agree that emotion is present and is negative (*irritated, annoyed,* etc; cf. figure 14). That no-clarification and first clarifications are likely to have listeners disagree as to emotional content is interesting; arguably, S1's speech is expected not to be emotionally charged early in the

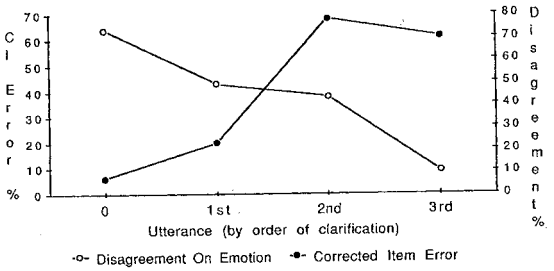

Figure 9. Judgments for emotion and contrastive emphasis as a function of the order of utterance in the dialogue.

dialogue, however, only some listeners perceive this.

Only five utterances were agreed by listeners to have no emotion on the utterance; all were no-correction utterances, where the intonational pattern on the digits is H* !H* !H* [5]. Vowel length of well-perceived, no-correction utterances judged to be non-emotional are given 10a, and those judged to be emotional are given in 10b. Comparing the difference between digits in 10a/10b in 10c/10d, respectively, shows that the relative length difference between the vowels in the three digit series is smaller in no-clarification utterances perceived as non-emotional, than in no-clarification utterances which listeners identify as emotional.

	initial	medial	final	I->M	M->F	I->F
a. $\bar{x}$	187	177	207			
b. $\bar{x}$	190	160	232			
c. $\bar{x}$				10	30	20
d. $\bar{x}$				30	72	42

Figure 10. Average vowel length and length differences for emotional and non-emotional, well-perceived, no-correction utterances.

Comparing F0 tracings of well-perceived no-correction utterances judged to be non-emotional (see figure 5) with no-correction utterances which some listeners judge emotional in figure 11 shows that utterances judged as non-emotional have a relatively stable intra-digit tonal contour, whereas those in 12 tend to be characterized by a broader F0 range, and greater F0 variation on the digits in the sequence [6].

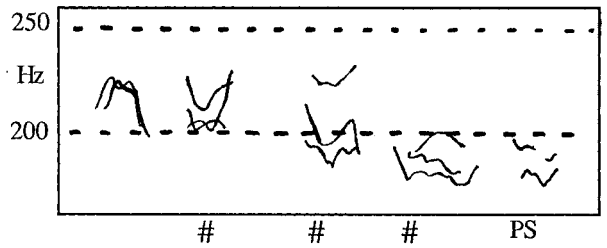

Figure 11. F0 tracings of emotional, no-clarification utterances.

Average vowel length of digits in initial-correction utterances which are poorly identified as initial-corrections, but judged to be emotional are given in figure 12. The initial digit averages 0 and 68ms longer than the medial and final digits, respectively, showing that the emotional modality does not entirely obscure the length of contrastive emphasis on the initial digit; it remains longer than the final digit, but not longer than the medial, and is shorter relative to the initial digit in well-perceived initial-correction utterances. There are no really poorly perceived final-correction utterances and the two fairly well-perceived (3-4/16 misjudgement rate) final-corrections were not agreed upon by listeners for emotional content.

	initial	medial	initial
$\bar{x}$	278	278	210

Figure 12. Average vowel length of digits in emotional, poorly perceived (6-15/16 misperception of CI) initial-correction utterances.

Figure 13 gives the emotional modalities assumed in this paper: words in capital letters indicate assumed modalities, and terms used by listeners, which we take to be instantiations of these modalities, are in small case letters; parenthesized numbers indicate the total number of times the term was used by all listeners in the entire study.

1. HAPPY-happy (11), happy-amused (1), happy-normal (2) TOTAL: (14); 2. AMUSED-amused (24), amused flustered (1), amused-frustrated (1), frustrated-amused (1), TOTAL:(27); 3. QUESTIONING-questioning (9), questioning-concerned (1), questioning-confused (1), uncertain-frustrated (1), surprised (3), TOTAL:(15); 4. NONE-none (19), normal (50), calm (54), normal-calm (1), TOTAL: (124); 5. CONCERNED-concerned (21), TOTAL: (21); 6. EMPHATIC-emphatic (103), emphatic-amused (1), emphatic-annoyed (1), urgent (7), urgent-annoyed (1), TOTAL:(113); 7. FLUSTERED-flustered (8) TOTAL:(8); 8.IRRITATED-irritated (48), annoyed (49), angry (3), frustrated (28), disgusted (3) TOTAL: (131)

Figure 13. Assumed emotional modalities based on terms used by listeners.

Four emotional modalities with broad qualitative differences are considered here: HAPPY, NONE, EMPHATIC, and ANGRY. Assuming the modalities in 13, the percent of emotional modalities identified by listeners correlates as we might expect with the order of the utterance in the dialogue: no-clarification utterances are judged to be most NEUTRAL, next HAPPY, third EMPHATIC, and seldom ANGRY. First clarifications are judged mostly EMPHATIC, then ANGRY, then NEUTRAL, and not HAPPY. Second clarifications are judged to be mostly ANGRY, then NEUTRAL, then EMPHATIC, and not HAPPY; and third clarifications are judged to be most ANGRY or EMPHATIC, sometimes NEUTRAL, and never HAPPY.

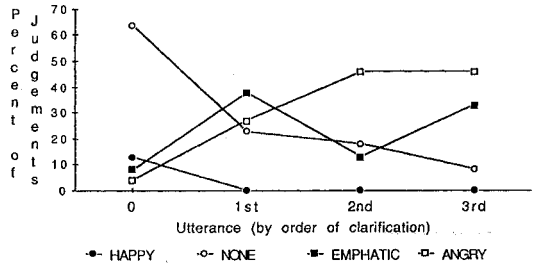

Figure 14. Listener judgement for four emotions by order of clarification utterance in dialogue.

5. SUMMARY

With the development of hierarchically organized linguistic representations, especially those denoting metrical structure [7] and intonation [1], [5], [8] as semi-autonomous tiers of information, the question of whether the representation of the non-lexical verbal expression of emotion is interactive with the abstract linguistic representation, or whether it is 'paralinguistic' is increasingly testable [3], [9], [10]. If emotional modality applies to the output of an intonated utterance then the metrical and intonational information of the morphosyntactic expression should, ceteris paribus, be invariant, with the verbal characteristics of emotion imposed on the output. If the verbal expression of emotion interacts with the intonation and/or metrical structure of the morphosyntactic representation, then the vowel length and F0 contours--indicating aspects of the metrical and intonational structure--of the morphosyntactic unit is expected to change as a function of the emotion expressed in the utterance. This pilot study addressed this question by examining the length and F0 correlates of contrastive emphasis in utterances varied for emotional modality. The results are insufficient to answer the question, but provide a step in the right direction.

Length was the most salient acoustic cue for the corrected digit in both initial and final-corrections, a finding which is consistent with those of Westbury and Fujimura [2]. Vowel length of the initial digit relative to the other digits is shortened in emotionally charged initial-correction utterances which are poorly perceived for the corrected digit; however, since poorly perceived final-corrections were virtually non-existent and the two extant fairly well-perceived final-correction utterances had listener disagreement for emotion, the question could not be answered for such utterances. The F0 correlate of contrastive emphasis reported in other studies [1] did not consistently occur on corrected items in these data and is genuinely rare in final-correction utterances. Even when the rising tone occurred--i.e on initial corrected digits--this cue was insufficient to unambiguously allow listener identification of the corrected item. Because of these findings, it is difficult to tell what aspect of the F0 contour, if any, consistently indicates contrastive emphasis and which indicates emotional modality in this study. Pierrehumbert and Hirschberg [1] characterize the intonation of correction as a L+H* pitch accent, but in the several cases examined here, where a rising tone occurred, it was characterized as as H* H- H%, rather than as a L+H* pitch accent. It appears that both of these rising contours can be employed by speakers to indicate contrast with an alternative item; possibly, the former, H*H-H%, is used on the corrected item when the item is in a sequence with overtly specified, or listed alternatives.

Corrected digits occurring later in the dialogue tended to be missed more, but the perception of emotion was better (assuming that listener agreement indicates speaker intent), and conversely, utterances early in the dialogue were well-perceived for corrected digit but disagreed upon for emotion. It is difficult to determine whether these two facts are related. It could be that when the utterance contains a significant amount of emotion, listeners miss the morphosyntactic information more. However, equally plausible is that utterances occurring later in the dialogue are perceived for emotion better because emotion is actually present, and corrected items are less well-identified in such utterances because the speaker literally fails to focus on the corrected digit as s/he becomes more emotionally involved.

REFERENCES

[1] J. Pierrehumbert and J. Hirschberg. "The Meaning of Intonational Contours in the Interpretation of Discourse." *Intentions in Communication*, P.R Cohen, J. Morgan, and M.E. Pollack, eds.1990.

[2] J. Westbury and O. Fujimura. "Articulatory correlates of contrastive emphasis in correcting answers in English." Presented at ATR Workshop on Speech Perception, Production, and Linguistic Structure. ATR, Kyoto, Japan. 1990.

[3] K. R. Scherer, D. R. Ladd and K. E. A. Silverman. "Vocal cues to speaker affect: Testing two models." *Journal of the Acoustical Society of America*, 76 (5), pp. 1346-1356. November, 1984.

[4] D. Erickson and O. Fujimura. "Acoustic and Articulatory Correlates of Contrastive Emphasis in Repeated Corrections." *ICSLP-92*, 1992.

[5] K. Silverman et al. Paper on TOBI/TRAX transcription system to be presented at *ICSLP-92*, Banff, Canada. October, 1992.

[6] C. E. Williams and K. N. Stevens. "Emotions and Speech: Some Acoustical Correlates." *JASA* (52) 4, pp. 1238-1249. 1972.

[7] M. Liberman and A. Prince. "On stress and rhythm." *Linguistic Inquiry* 8, pp. 249-336. 1977.

[8] M. Beckman and J. Pierrehumbert. "Intonational structure in Japanese and English." *Phonology Yearbook 3*, pp. 255-309. 1986.

[9] K. D. Emmorey. "The Neurological Substrates for Prosodic Aspects of Speech." *Brain and Language* 30, pp. 305-320. 1987.

[10] K. R. Scherer. "Vocal Correlates of Emotional Arousal and Affective Disturbance," *Handbook of Social Psychophysiology*. H. Wagner and A. Manstead eds. 1989.