Computer Speech and Language (1994) **8**, 385–404

Quantitative description and differentiation of fundamental frequency contours

Christopher A. Moore,*‡ Jeffrey F. Cohn† and Gary S. Katz†

**Division of Communication Science and Disorders and †Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.*

Abstract

Fundamental frequency (f_0) contours derived from the speech of 35 mothers to their 4-month-old infants were quantified for two experimental conditions, one in which the mother was instructed to seek her infant's attention and a second in which the mother was instructed to express approval of her infant's action. In addition to conventional descriptions (e.g. mean and standard deviation of f_0, and utterance duration) the contours were subjected to modelling using 16 equations (1 linear and 15 non-linear) selected to reflect customary qualitative descriptions of f_0 contours (e.g. Gaussian, rising, falling). Curve-fitting results confirmed that these infant-directed utterances were fit extremely well by at least one function. The average maximum R^2 value obtained across all 16 equations was 0·83. Furthermore, discriminant analysis demonstrated that these two utterance types could be differentiated with 76% accuracy by these curve-fit results alone. Discrimination improved to 92% accuracy, however, when additional, more global descriptors were included in the discrimination function (e.g. utterance duration, mean fundamental). These results suggest that infants may be responsive to specific prosodic stimuli, which may involve distinct voice dynamics or more general speech signal characteristics, such as overall pitch, pitch variability, or utterance duration.

1. Introduction

The suprasegmental dimensions of speech, which are usually conceptualized as being overlaid on lexical, syntactic, and semantic linguistic constructs, have been widely acknowledged as carrying a significant portion of the communicative intent of the speaker. This information can be supplemental to that provided by the linguistic structure of the utterance, or can provide essential elements of the intended meaning. For example, in one commonly used experimental paradigm, it is only suprasegmental, or prosodic, information that provides the contrastive stress necessary to distinguish the utterance "BEV loves Bob?", in which the speaker is questioning his assumption that someone else loves Bob, from "Bev LOVES Bob?", in which the speaker is

‡To whom correspondence should be addressed at: 3347 Forbes Avenue, University of Pittsburgh, Pittsburgh, PA 15260.

0885–2308/94/040385+20 $08.00/0

questioning his assumption that Bev does not love Bob (e.g. Cooper & Sorenson, 1981). In addition to their "syntactic" function, suprasegmentals convey essential information about a speaker's expressed emotion. Elation, for instance, is conveyed by increased mean f_0 and f_0 range (Scherer, 1986). Nevertheless, despite ready acknowledgment of the importance of suprasegmentals, researchers have been challenged to develop a quantitative descriptive framework for this essential level of spoken language (Frick, 1985).

There are two obstacles to an adequate quantitative description of prosodic structure, the more obvious of which is the lack of a sufficiently precise definition (i.e. one that is both exhaustive in terms of the contributing phenomena and exclusive in terms of other linguistic and paralinguistic dimensions) of what physical parameters comprise prosody. Modulations of loudness, vocal pitch and vowel quality are usually included in descriptions of prosody, although, among these parameters, vocal pitch might be legitimately represented as primary, especially to the extent that loudness and pitch covary. The second problem characteristic of prosodic parameters (e.g. vocal pitch and loudness) is that their variation is not rigidly specified by linguistic rules, leaving them highly susceptible to large inter- and intraspeaker variations, and to variations stemming from linguistic content. The multitude of factors affecting these parameters may preclude extracting the essence of the prosodic structure of an utterance, but it seems reasonable to seek to isolate and specify individual prosodic parameters, such as variation in vocal pitch in order to evaluate their contribution to spoken communication. The present investigation was designed to address this second problem of describing prosodic consistencies, of which we are well aware as speakers and listeners, but find difficult to describe in quantitative terms.

It is possible to create experimental conditions or to observe natural conditions in which these difficulties are considerably reduced. Speaking contexts characterized by accentuated and stereotypic prosody may yield wider ranging utterances with proportionately less variability across repetitions. Recitation of rote phrases (e.g. nursery rhymes, the alphabet) and speech intended for listeners with reduced perceptual capacities (e.g. infants, pets) characteristically exhibit these properties (Ferguson, 1964). One promising direction of investigation has been the investigation of infant-directed speech. Speech directed to infants is characterized by exaggerated pitch and intensity changes and tends to exhibit a great deal of stereotypy (Stern, Spieker & MacKain, 1982; Stern, Spieker, Barnett & MacKain, 1983; Fernald & Simon, 1984; Fernald, 1989). Moreover speakers intuitively recognize that their message is primarily, if not solely, communicated via suprasegmentals.

Appropriately taking as justification the essential contribution of prosodic features to spoken communication and the need to communicate their description into more general models of language, most researchers have had to rely on qualitative observations of vocal pitch and intensity contours to describe this important level of communication (e.g. Garnica, 1977; Stern, Spieker & MacKain, 1982; Stern *et al.*, 1983; Fernald & Simon, 1984; Bettes, 1988; Fernald, 1989). Fernald and her colleagues (Fernald & Simon, 1984; Fernald, 1989), and Stern, Spieker, and MacKain (1982) have employed such descriptors: rising, bell-shaped (or rise-fall), slow falling, rapidly falling, and complex. Operational definitions of these terms (e.g. inclusion of only those f_0 contours with excursions of more than 128 Hz (Stern, Spieker & MacKain, 1982) or more than 13 semitones/s (Fernald & Simon, 1984), have enhanced quantifiable descriptions, but have stopped well short of quantification of the shapes themselves. Notable exceptions

have extracted target prosodic dimensions from running speech and applied those to quantifiable correlates of the intended message (e.g. Pierrehumbert, 1990; Price, 1991). The unfortunate consequence of this necessary accommodation of these especially difficult data has been twofold: subjective interpretations are based on logistically limited numbers of observations (i.e. reliable categorization is time-intensive), and descriptions emerge that force naturally occurring utterances into more or less contrived categories based upon an investigator's *a priori* assumptions of the signal's modulation.

What is required then is a method by which the essential suprasegmental properties of an utterance can be extracted, quantified and summarized. The present investigation was designed to address this need by evaluation of the discriminative potential of a quantitative approach for one primary prosodic element f_0, under conditions created to elicit exaggerated exemplars. Vocal fundamental frequency has several desirable characteristics as a dependent variable related to linguistic and affective intent: f_0 is one of the primary contributors to prosodic structure, is relatively easy to extract from an acoustic signal, extends over the duration of a phrase or utterance, and has well known underlying control parameters (e.g. vocal fold length and tension). Specifically we focused on modulation of vocal fundamental frequency during mothers' speech directed to their 4-month-old infants under conditions in which intended meaning was controlled. Given an infant's putative lack of lexical, syntactic, or pragmatic knowledge, we assumed that, whatever the message communicated during infant-directed speech (IDS), it must be carried predominantly by suprasegmental features. (Recent work by Mandel and Jusczyk, 1994, however, has demonstrated lexical processing in 4·5-month-old infants presented with their own names as stimuli.) We exploited this limitation by directing the mothers to convey specific messages, without specifying lexical content, of approval or attention-seeking to their infants.

2. Methods

2.1. Subjects

Subjects of this investigation were the mothers from 25 mother–infant dyads. The age of their infants at the time of the observations was 4 months (SD: 1 week). Mothers were screened for depression, which has been shown to affect IDS (Breznitz & Sherman, 1987; Bettes, 1988). All infants were first-born and experienced normal, full-term births.

2.2. Experimental protocol and conditions

Each mother–infant pair was seated face-to-face in a sound-treated studio. Following a brief period of orientation and instruction there were two ordered conditions, approximately 30 s to 1 min each in duration, during which the mother communicated two different specified messages to her child. These two interaction conditions were intended to elicit distinct pitch contours and were assumed to represent normal communication conditions between mothers and their infants. The two conditions included:

- an attention condition (AT), in which the mother was directed to draw her baby's attention to a red ring;

- an approval condition (AP), in which the mother was directed to express approval of her baby's touching red ring.

Mothers reported that they were comfortable with these requirements and that they felt that their productions were accurate representations of how they normally speak to their babies.

Our selections of attention-eliciting and approval conditions were intended to meet several criteria. These conditions represent commonly observed verbal behaviours in mothers of young infants (Fernald & Simon, 1984), have been studied previously in investigations of IDS (Fernald, 1989), and have been shown to be characterized by contrasting pitch contours (Stern, Spieker & MacKain, 1982; Fernald & Simon, 1984; Fernald, 1989). Specifically, attention-eliciting utterances are characterized by a rising pitch contour, and expressions of approval are typically observed to manifest a rhythmically rising and falling contour, which has been referred to as "sinusoidal" in shape (Stern, Spieker & MacKain, 1982; Fernald, 1989).

2.3. Data collection

Data sources included:

- split-screen video recording of the mother's and infant's faces viewed face-on;
- audio recordings obtained from a microphone worn by the mother;
- accelerometric recordings obtained from accelerometers worn by the mother over the thyroid cartilage of her larynx. These records provided isolated signals derived from vibration of the larynx associated with vocal fold movement, and yielded high quality waveforms suitable for fundamental frequency (f_0) extraction;[1]
- both video channels, of all three audio/accelerometer channels, and a SMPTE (Society of Motion Picture and Television Engineers) digital time code channel were recorded directly to a professional quality videocassette recorder for subsequent analysis. Only the acceleration signal was used in the present analysis, although the remaining channels were used to clarify the location of the stimuli on the tape and confirm the compliance of the mothers with the experimental protocol.

2.4. Signal processing and initial data reduction

Our goal was to obtain one utterance for each of the two conditions from each mother to comprise a final dataset of 50 utterances. Because of occasional problems in signal quality (e.g. production of high amplitude utterances by the mother, which overdrove the amplification systems) and varying degrees of success by the mothers in complying with the experimental directions, mothers contributed a variable number of utterances to the initial dataset from which the final set was randomly selected. On average, each mother produced about three acceptable utterances per condition (range: 1–8 utterances). In all, 364 utterances were recorded, of which 42 were discarded, 36 because they had less than the minimum of 30 data points required for subsequent analyses, and six because they yielded contours judged to be too inconsistent to be subjected to the semi-

[1] In fact, microphones or accelerometer-based signals could be used effectively. Statistical comparison of microphone vs. accelerometer-based signals by condition revealed no significant difference in the source of the fundamental frequency signal [MANOVA $F(10,39)=0{\cdot}946$; $P=0{\cdot}504$].

automated analysis. Inclusion criteria for utterances in the initial dataset also included a judgment by naive research assistants, who judged each utterance with respect to the mother's success in matching the constraints of the condition (i.e. expressions of approval or elicitation of attention).

Initial processing of the recorded samples required two stages of parsing from the taped sessions down to the individual utterances for analysis. First a segment of each condition was generated by two trained observers. This segment exactly identified the beginning and end times (using the SMPTE time code on the tape) of each experimental period. The complete experimental periods were digitized using a commercially available speech analysis hardware/software system (CSL, Computerized Speech Laboratory, Model 4300, Kay Elemetrics Corp.) at 5000 samples/s to yield two large digitized records for each mother, one for each experimental condition. The individual utterances analysed were subsequently parsed from these files. Trained observers parsed individual utterances from each condition on the basis of two operational criteria defining an utterance: (1) an utterance represented the minimum isolated string that could stand alone as a complete phrase or sentence (e.g. "Good boy! That's good! Good boy!" was parsed into three utterances); and (2) the utterance met the prescribed task conditions of approval or attention eliciting (per Papousek, Papousek & Bornstein, 1985).

Internal validity of token parsing was assessed by a second parsing of the digitized samples by a second coder. A randomly selected sample representing 21% of the total dataset was used. Within each sample, an agreement was counted when the two coders agreed that one, and only one, token occurred in the same segment of speech. Disagreements were defined as those segments identified by one coder, but not by the other. Interobserver agreement was 86 and 72% for identification of approval and attention-eliciting utterances, respectively. Because rejected tokens (i.e. those periods rejected as utterances by both observers) were not tallied, these values must be interpreted as significant underestimates of the observers' performance reliability.

The final dataset was comprised of one utterance randomly selected from each condition for each mother. The pitch contour for each utterance was extracted using a computer-assisted implementation of the CSL pitch extraction algorithm. Because speech signals with higher fundamental frequencies are much more difficult to analyse than those with a lower f_0 (as seen in adult male speakers), this algorithm was designed to simultaneously provide a large number of analysis options for each utterance, thereby allowing the technician to quickly select the optimal parameters for each speaker and condition. For each trace, the narrow band spectrogram and up to eight different f_0 contours were computed and displayed by varying the width and overlap of the analysis window applied to the digitized speech sample. Taking the shape of the narrow band spectrogram as the most faithful representation of f_0 modulation, the technician selected the f_0 contour that best matched the contour suggested by the narrow band spectrogram. The numerical values and time markers of the selected contour were then exported to an external file for subsequent processing and modelling.

Raw pitch contours extracted using the CSL exhibited characteristic "dropouts" (i.e. isolated analysis frames in which no pitch was detected even though they are bounded by continuous f_0 contours) and gaps (i.e. periods of silence or non-voiced speech sounds) typical of most pitch extraction algorithms. These discontinuities were unacceptable for submission to mathematical modelling, which made additional processing necessary. The curve-fitting algorithm accommodated non-voiced intervals by ignoring zero values, as these points were not seen as contributing to the overall perception of the pitch

contour and because sudden changes in value yielded very poor fits. This restriction required correction and smoothing of the raw f_0 contours. A customized program was used to evaluate the variance within a moving rectangular window three points wide and corrected pitch estimation errors by doubling or halving the value of each successive point to minimize the variance within the window (Katz, Moore & Cohn, 1992). The assumption underlying this process was that most f_0 estimation errors result from recognition of the wrong periodic signal as the fundamental. In fact the erroneous estimates were usually half or twice the actual f_0 (i.e. $5f_0$ or $2f_0$), so that whenever the variance computed by the algorithm could be minimized by doubling the last point in the window, an error of exactly half the real pitch (e.g. as if the algorithm had failed to detect one zero crossing, yielding a frequency half of the real frequency) was assumed and this point was adjusted accordingly. The technician inspected the result of this correction to again verify the match of the extracted contour to the spectrographic representation of the utterance. Following this correction of dropouts and pitch extraction errors, these contours were smoothed using a commercial, Fast Fourier Transform (FFT) based smoothing algorithm (TableCurve, Jandel Scientific). These smoothed functions provided the final dataset for modelling. Fig. 1 illustrates the three stages of this process of correction and smoothing.

The three panels of Fig. 1 demonstrate the progression from the initial f_0 contour extracted using optimized parameters in CSL (a) through the corrected (b) and smoothed (c) signals. The abrupt drops from the main contour are half-frequency errors resulting from the low spectral density of higher frequency voices. These errors were corrected by doubling the extracted frequency value to achieve the minimum variance among adjacent points. In the example provided, the mean fundamental frequency can be seen to be approximately 350 Hz. The lower spectral density resulting from the widely spaced harmonics of the glottal source provides a relatively poor input signal for f_0 extraction, with the most common error being one of extracting a frequency value that is half the actual value (e.g. in a simple algorithm based on counting zero crossings, this would be the equivalent of missing one of the zero crossings). Although the original contour might have been acceptable for many applications, modelling required elimination of discontinuities and compensation for quantization error, which is characteristic of f_0 contours. Finally, a conventional FFT-based smoothing algorithm was applied to reduce the irregularities in the data caused by quantization error, the error introduced by the relatively gross frequency resolution possible for brief non-stationary signals.

2.5. Modelling of pitch contours

The corrected and smoothed f_0 contours were subjected to modelling using a selected subset of functions available in a commercial curve-fitting package (TableCurve, version 3·2, Jandel Scientific). We selected this subset of functions on the basis of their ecological validity, as evidenced by their natural occurrence in biological systems, and because they correspond to the qualitative descriptors used in previous research [i.e. rise, fall, bell-shaped, and sinusoidal or complex (Stern *et al.*, 1982; Fernald, 1989)]. There were 16 functions grouped into seven classifications of curve type. (Classifications were exclusive except for the power function, which was analysed in isolation as its own classification and also as a member of the exponential classification.) Examples of each of the classes of equations are provided in Fig. 2, which includes both rising and falling variants of each function where appropriate. The 16 equations and their classifications

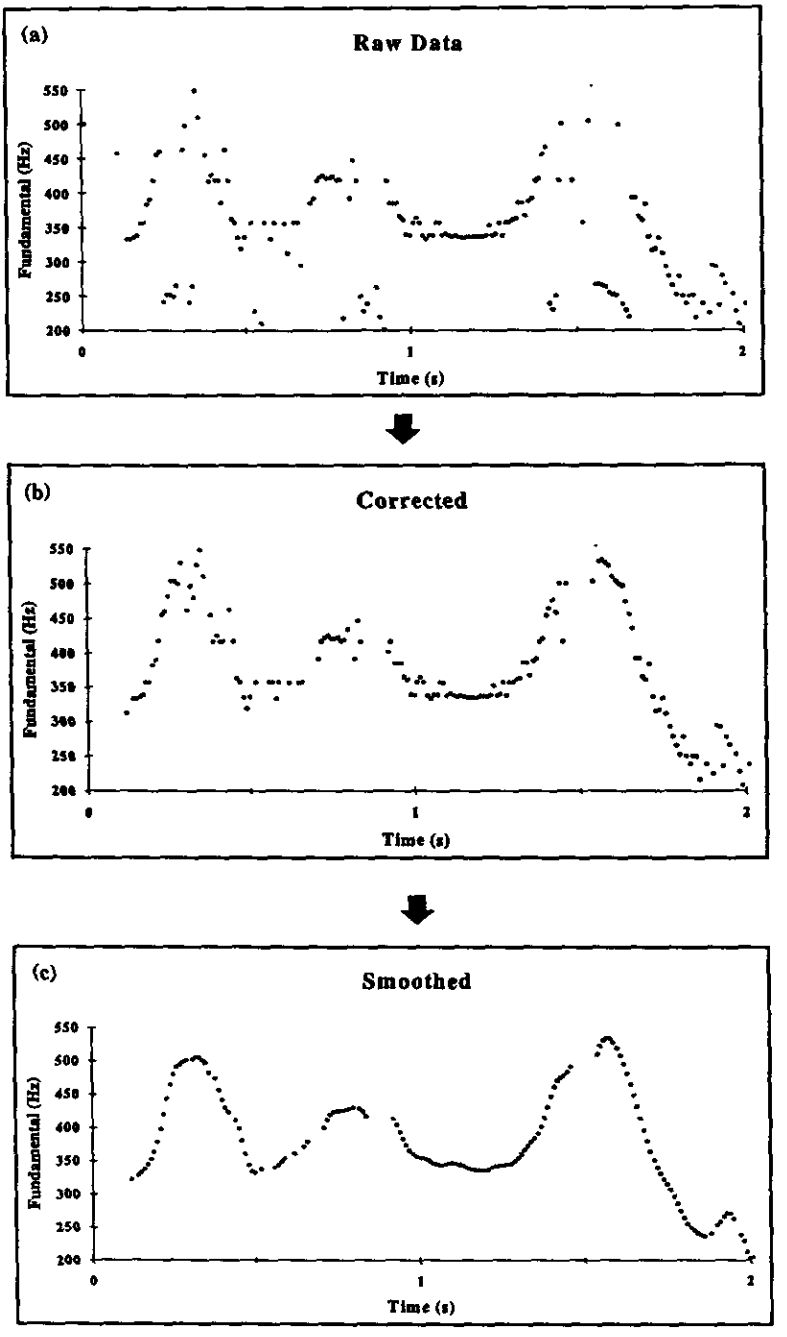

Figure 1. Illustration of initial signal processing following f_0 extraction. The initial 2 s contour (a) exhibits frequent discontinuities, which result in part from the f_0 extraction algorithm. Many of these discontinuities could be resolved by correction for half-frequency errors (b). Quantization error, the result of analysing fixed length frames of digitally represented waveforms, was reduced by Fast Fourier Transform smoothing to yield the signal subjected to modelling (c).

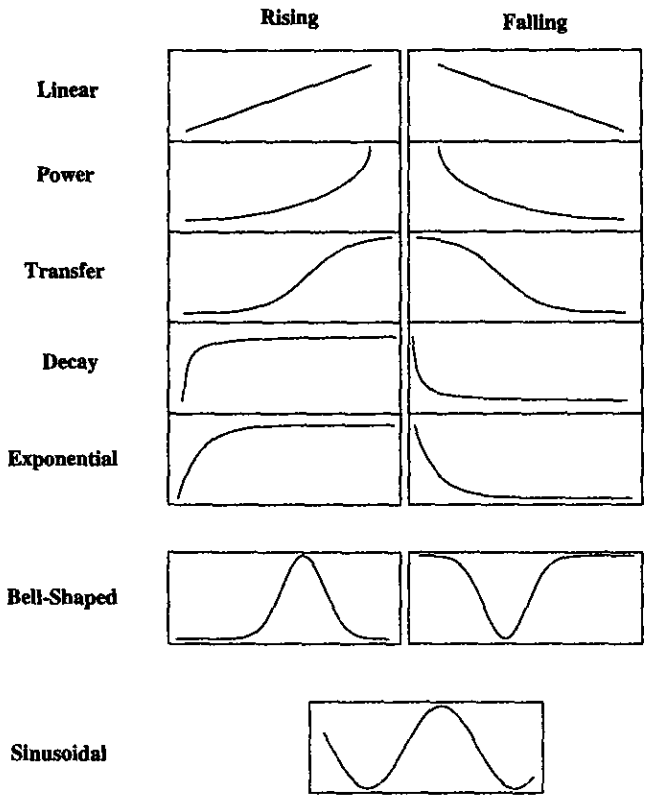

Figure 2. Idealized representation of the seven types of functions applied to f_0 contours. Rising and falling variants are shown where applicable. Each category consisted of from one to six specific functions.

are provided in the Appendix. The smoothed f_0 contours were submitted for curve fit analysis to each of these 16 functions. Goodness of fit, expressed as R^2, the proportion of variance in the token explained by the model, was evaluated for each function and was also pooled within each of the seven classifications to yield an index for each classification (e.g. a peak index represented the average R^2 value of the six functions in that classification). These indices provided a further reduced dataset upon which subsequent statistical analyses were based.

2.6. Additional descriptors of fundamental frequency contours

In addition to the curve-fitting analysis, several simple descriptors were computed for each utterance. These parameters were selected as being representative of those most frequently used in analyses of fundamental frequency and provided a comparative base for the curve-fitting analysis. These conventional descriptors included: mean, standard deviation, range and duration of the f_0 contour. These values were computed automatically for each corrected and smoothed f_0 contour, and were associated with the fit indices of the modelling analysis to comprise the complete dataset, which was subjected to discriminant analysis.

These two different parts of the analysis, curve-fitting and simple description of more global quantities, represent two very different approaches to the data. Analysis of the fit of an f_0 contour to a specific curve type entails the assumption that it is the shape of the contour that distinguishes utterance types (e.g. a Gaussian-shaped contour would be assumed to carry a meaning distinct from a decay contour). On the other hand, descriptive statistics of the entire utterance imply that differentiation of meaning is based on parameters such as the overall pitch of the utterance, the extent of the pitch range covered, or simply the length of the utterance. These two approaches are not mutually exclusive, however, and it seems likely that there is covariation among some factors. For example, more complex contours (e.g. sine functions) probably take longer to produce, such that the fit for sinusoidal functions covaries with duration.

3. Results and discussion

3.1. Modelling of f_0 contours

Our efforts to model these naturally occurring f_0 contours were generally quite successful. Considering the fit of all 16 functions to a given f_0 contour, the average maximum R^2 value was 0·83 (SD = 0·17; range 0·13–0·99) for all 50 utterances, confirming that in most cases at least one finite function accurately described the modulation of fundamental frequency during an utterance (i.e. accounting for 83% of the variance on average). The average maximum R^2 values for the attention and approval conditions were 0·85 (SD = 0·15; range = 0·42–0·99) and 0·82 (SD = 0·20; range = 0·13–0·99). These coefficients were not significantly different [$F(1,48) = 0{\cdot}844$; $P = 0{\cdot}36$]. This finding supports the suggestion that the modelling success of these techniques does not vary with utterance type or suprasegmental pattern.

Fig. 3 illustrates the results of the algorithm in modelling two utterances using two of the 16 equations, the sine function and the Gaussian peak function. The utterances in this figure demonstrate excellent differentiation based on curve-fitting results with each f_0 contour being fit well by only one of the two functions shown. Fig. 3(a) includes the fit of a sine function to an attention utterance. The R^2 value for the fit of this curve to the f_0 data was 0·108 ($R^2_{adjusted} = 0{\cdot}000$). Fig. 3(b) illustrates the fit of the Gaussian peak function for the same pitch contour; R^2 was 0·875 ($R^2_{adjusted} = 0{\cdot}858$) in this case. Fig. 3(c) and (d) illustrate the fit of the same two functions to an approval utterance. In these cases the sine function seen in (c) yields a much better fit ($R^2 = 0{\cdot}875$, $R^2_{adjusted} = 0{\cdot}870$) than the Gaussian peak function seen in (d) for the same approval utterance ($R^2 = 0{\cdot}062$, $R^2_{adjusted} = 0{\cdot}022$). It is evident from this figure that very subtle differences in f_0 contour can give rise to significant modelling differences. This level of sensitivity is very desirable in modelling attempts and in explanations of prosodic phenomena, because it may be these small variations in vocal dynamics that yield perceptual differences, which may be essential to transmission of intended meanings.

Two questions arise with respect to the success of the different functions and functional categories in fitting specific functions: How well can mathematical functions describe f_0 contours? and, Can these functions differentiate among intended meanings? Fig. 4 illustrates the distribution of maximum coefficients for the complete set of equations. Separate distributions for the approval and attention conditions are given, as well as the combined distribution. From this figure it is clear that the overall performance of the peak function category contained the largest number of best fits, not only because

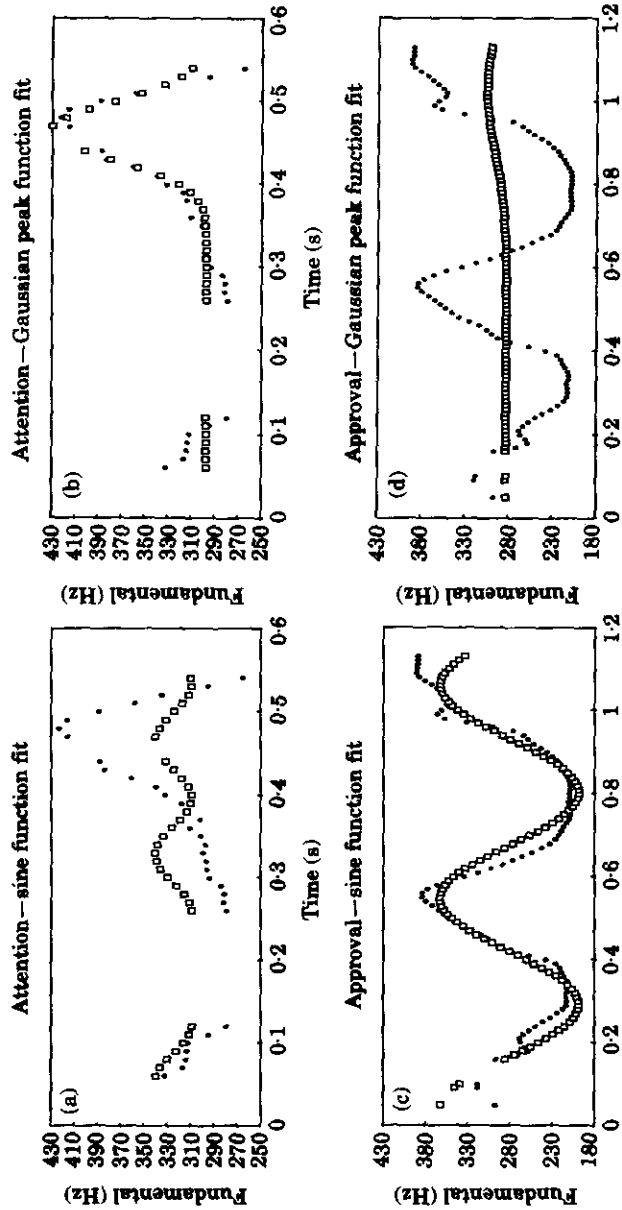

Figure 3. The results of modelling *two* f_0 *contours, one* an attention-seeking utterance, which was best fit by a sine function, and the second, an approval utterance, which was best fit by a Gaussian function. Actual f_0 contours are shown as solid lines, the best fit for sine and Gaussian functions are shown as dotted lines.

it is the largest category (i.e. with five functions), but also because the three functions that most frequently fit the contours best (extreme value, error function and Gaussian peak functions) were all members of this category. The intuitive appeal of this finding was that subjective impression of these contours was that they usually tended to rise and fall, although the height and depth of each phase was quite variable. Because the specific curve fit to each contour was for a finite set of points, this finding does not suggest that every contour rose from and fell back to a baseline level. Rather, the finding was taken to suggest that these contours frequently reached a peak prior to the end of the utterance. Furthermore, though it was mathematically possible for these functions to demonstrate negative- rather than positive-going peaks, there were no utterances that were best fit by negative-going peak functions. These results demonstrate very consistent success in simple description of f_0 contours using a small, fixed set of ecologically valid mathematical functions.

3.2. Discrimination of utterance types

The second question of interest in this investigation was whether or not utterance types could be discriminated on the basis of differential curve-fitting results. Fig. 3 does not help to resolve this issue, although there are some obvious differences in the results for the two utterance types. For example the error function peak equation yielded the best fit for attention utterances much more frequently than for approval utterances, and the results for f_0 contours fit to the sine function showed the opposite effect. Accordingly, evaluation by discriminant function analysis was completed to assess the overall utility of these measures as discriminators, as well as the specific discriminative value of the selected functions. To avoid overspecification of the model derived using discriminant function analysis, we aggregated the curve-fit results within each function type, with each resulting aggregate representing the average of one to six R^2 values obtained for individual functions. Aggregation reduced the number of parameters from 16 equations to seven equation types.

The reliability of the composition of the function classification groups with more than one equation (i.e. exponential, peak, waveform and transition functions) was assessed by reliability analysis of each group. Table I summarizes the results of these reliability analyses. All α coefficients were highly significant, confirming the assumption that the functions within each group were each contributing to the group index, and that the average of the R^2 values yielded a more reliable measure than any of the individual coefficients. In fact, there was only one function for which removal of that function from its classification group increased the group α coefficient, that being logistic dose-response transition function in the transition group. In that case, removal of the function yielded a change in the α coefficient from 0·840 to 0·911. Nevertheless, there being no *a priori* justification for removing this single function, it was retained in all subsequent analyses. This finding supported the aggregation of results for individual functions into types before submitting these data to discriminant function analysis.

Several additional preliminary steps were required to elucidate the context and the limits of the discriminant function analysis. Initially we evaluated whether or not there were differences in success of different equation types in modelling these utterances, and whether or not there were significant differences between utterance types within any of the individual function types. Table II summarizes the results of curve fitting

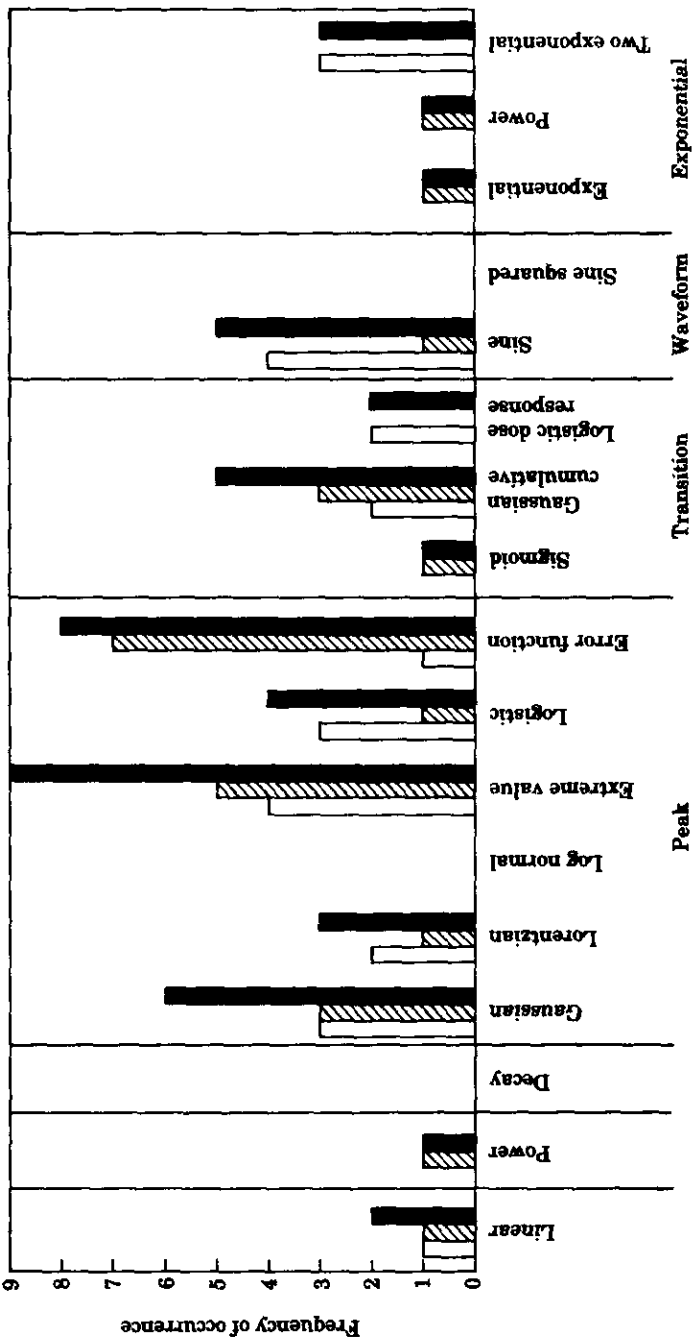

Figure 4. The distribution of best fit (i.e. the function for which the highest R^2 was obtained for each utterance) across the 16 equations and seven equation types for all 50 utterances. (□), Approvals; (▨), attention-seeking utterances; (■), totals for both categories.

TABLE I. Reliability analyses for composition of classification groups with multiple functions contributing to the group index

Classification group	Number of functions in group	α Coefficient
Peak	6	0·942
Transition	3	0·889
Waveform	2	0·995
Exponential	3	0·598

TABLE II. Means and standard deviations of R^2 indices (i.e. R^2 values aggregated within function classifications) across subjects

Utterance type	Function classification						
	Linear	Decay	Peak	Transition	Waveform	Exponential*	Power
Approval	0·31 (0·30)	0·13 (0·22)	0·66 (0·26)	0·45 (0·37)	0·64 (0·29)	0·26 (0·29)	0·38 (0·32)
Attention	0·17 (0·29)	0·16 (0·20)	0·77 (0·19)	0·54 (0·38)	0·57 (0·32)	0·43 (0·30)	0·53 (0·34)
Combined	0·24 (0·30)	0·14 (0·21)	0·72 (0·23)	0·49 (0·38)	0·61 (0·30)	0·34 (0·30)	0·45 (0·34)

* Between group pairwise comparison, $F(1,48)$, $P<0·10$. For all other comparisons, $P>0·10$.

for each equation type (i.e. 16 equations of seven types applied to all 50 utterances, then aggregated by equation type). Inspection of this table readily reveals the striking findings that peak and waveform functions generally were able to fit these contours with the least error (average $R^2=0·72$ and 0·61, respectively), and the fits to the decay function were especially poor ($R^2=0·14$). Furthermore, statistical comparison of curve-fit indices between utterance types yielded only one significant difference ($P<0·10$); exponential functions fit attention utterances (average $R^2=0·46$) better than approval utterances (average $R^2=0·30$). What is not revealed in this table, however, is whether or not some linear combination of the results for each index might provide an accurate predictor of utterance type. Discriminant function analysis was undertaken to address this question. As an aside, it is important to note, with respect to the bulk of the extant literature in prosodic contours, that the finding for the linear index of an average R^2 of 0·35 suggests that representation of f_0 contours as "rising" or "falling" is, at best, inaccurate, especially to the extent that linearity is implied by these terms (cf. Fernald & Simon, 1984; Fernald, 1989; Papousek, Papousek & Symmes, 1991).

The correlated variation of R^2 indices was also evaluated. In order to qualify as good candidates for discriminating between utterance types, it was essential that the indices not be highly correlated across utterance types. Table III includes the correlation matrix resulting from a correlational analysis of all possible pairs of indices. The mean r^2 value, 0·286 (SD = 0·125; range 0·002–0·559) indicated that indices were only weakly

TABLE III. Correlation coefficients (r^2) among function categories for all f_0 contours

Function category	Linear	Decay	Peak	Transition	Waveform	Exponential	Power
Linear	1·000						
Decay	0·337	1·000					
Peak	0·201	0·192	1·000				
Transition	0·328	0·233	0·240	1·000			
Waveform	0·396	0·002	0·306	0·222	1·000		
Exponential	0·355	0·217	0·129	0·477	0·168	1·000	
Power	0·421	0·306	0·384	0·559	0·235	0·295	1·000

TABLE IV. Within utterance means and standard deviations of f_0 and utterance duration

Utterance type	Mean f_0 (Hz)*	Standard deviation of f_0 (Hz)†	Duration (s)‡
Approval	345·73 (81·97)	87·21 (37·67)	1·05 (0·37)
Attention	295·31 (103·89)	47·01 (21·89)	0·57 (0·27)
Combined	320·52 (96·05)	67·11 (36·63)	0·81 (0·40)

* $F(1,48) = 3·63$; $P<0·10$;
† $F(1,48) = 21·28$; $P<0·0001$;
‡ $F(1,48) = 28·52$; $P<0·0001$.

correlated across utterance types. This finding increased the discriminative potential of the modelling indices.

Another consideration in differentiating these utterance types was a comparison of modeling with differentiation and description using more conventional descriptors, which are generally descriptive statistics derived from the global utterance. The major weakness of the descriptors obtained in the present investigation (i.e. f_0 mean, standard deviation and duration) is that they fail to provide an accurate description of the f_0 contour, although they may well serve simply to differentiate among different utterance types. Accordingly, it is important to evaluate discrimination by modeling with reference to what is possible through more conventional means.

A summary of the global descriptors obtained for each utterance type is provided in Table IV. The differences between utterance types among these descriptors are quite evident. Approval utterances had higher mean f_0 values [$F(1,48) = 3·63$, $P<0·10$], exhibited more variability in f_0 within each utterance [$F(1,48) = 2128$, $P<0·001$], and were longer in duration [$F(1,48) = 28·52$, $P<0·001$] than attention utterances. These criteria obviously differentiated the two utterance types, which might have been expected to be quite discriminable by these features, given our intuitive impressions of attention-seeking utterances and expressions of approval. Accordingly two discriminant function analyses were completed, one using only the curve-fit results, and a second in which the global descriptors were included.

The first discriminant function analysis successfully differentiated pitch contours of the two utterance types on the basis of curve-fit results alone. These results are shown

TABLE V. Stepwise results of discriminant function analysis for curve-fit indices

Step	Variable entered	*F*	*P*
1	Exponential index	3·770	0·058
2	Linear index	4·917	0·012
3	Power index	4·645	0·006

TABLE VI. Classification results of discrimination analysis using curve-fit indices

Actual group membership	Predicted group membership	
	Approval	Attention
Approval	19	6
Attention	6	19

$\chi^2 = 13{\cdot}52$; $P < 0{\cdot}001$;
$\kappa = 0{\cdot}52$; $P < 0{\cdot}001$.

in Table V. Of the seven equation types evaluated, only three entered into the discriminant function. Predictably, given the between group difference for curve fitting, the exponential index entered the equation first, followed by the linear and power indices. As seen in Table VI this analysis yielded correct classification of 76% of the test utterances, with correct and incorrect identifications distributed equally between types. This finding, though not especially strong, confirmed that modelling of f_0 contours using ecologically valid mathematical functions is an effective approach to differentiating a single pair of utterance types. The more important aspect of this discrimination is that its derivative steps include an accurate mathematical description (the average maximum R^2 value across all 16 functions for each utterance was 0·83) of each f_0 contour. It is this combined result of discrimination and description that we view as essential to the study of the large datasets needed to form an understanding of prosodic phenomena.

The second discriminant function analysis, which included the three global descriptors as well as the curve-fitting results, yielded a six factor model including four curve-fit indices and two global variables. The results are shown in Table VII. Classification by this model, shown in Table VIII, yielded 92% accuracy with three erroneous classifications of approval utterances and only one for attention utterances. Although this is a very strong finding with respect to our ability to classify naturally occurring f_0 contours, it

TABLE VII. Stepwise results of discriminant function analysis including curve-fit indices and global measures (mean f_0, standard deviation of f_0, and duration)

Step	Variable entered	*F*	*P*
1	Duration	28·520	0·000
2	Standard deviation of F_0	23·207	0·000
3	Peak index	17·800	0·000
4	Linear index	14·556	0·000
5	Exponential index	13·576	0·000
6	Power index	12·107	0·000

TABLE VIII. Classification results of discrimination analysis when curve-fit indices and global measures are included

Actual group membership	Predicted group membership	
	Approval	*Attention*
Approval	22	3
Attention	1	24

$\chi^2 = 35{\cdot}57$; $P<0{\cdot}001$
$\kappa = 0{\cdot}84$; $P<0{\cdot}001$

is important to recognize that this classification was completed using only two utterance types under exaggerated speech conditions (i.e. infant-directed speech).

4. Summary and conclusions

The results of the present investigation demonstrated the successful extraction, modelling and discrimination of f_0 contours obtained in two distinct prosodic conditions, expressions of approval and attention-seeking utterances by mothers to their 4-month-old infants. The primary finding that we can model naturally occurring f_0 contours using a small set of ecologically valid functions is especially significant as it provides one of the essential tools for quantification of large sets of prosodic data (as noted by Scherer, 1986). Such a descriptive tool has been recognized as a vital precursor to furthering the study of speech prosody. The classification of suprasegmental properties into a closed set of known models has important ramifications beyond the study of prosody. It might be possible, for example, to synthesize, store, or transmit natural speech in separately analysed parts, one of which would be suprasegmental shape (others might include initial source spectra, onset and offset points, changing filter functions). Given a closed set of models, an f_0 contour can be accurately specified given only the equation, its limits and the coefficients. An additional finding is this descriptive technique can be used to discriminate different types of contours; but this remains

to be tested more rigorously using larger datasets obtained under less exaggerated conditions.

Several important considerations emerge to guide future investigation. Development of a thorough model of speech communication must not only yield accurate discrimination of utterance types, but must faithfully represent the components of the process as contributed and limited by both the speaker and listener. For example, the importance of global utterance descriptors, such as mean and variability of f_0 must be approached from both the listener's and the speaker's perspective. It is possible that the dominant psychoacoustic factors in determining the listener's perception are relatively gross factors such as the speaker's overall loudness and pitch, which may correlate well with levels of arousal (Scherer, 1986). Alternatively the dynamics of suprasegmental features, such as rate or "acceleration" (d^2f_0/dt^2) of pitch and loudness change may be perceived with enhanced sensitivity. Indeed the sensitivity of humans to change and differences rather than static conditions makes these factors especially attractive as carriers of suprasegmental information.

Finally it must be recognized that speakers and listeners are sensitive to more than static or dynamic suprasegmental parameters. Even at 4 or 5 months of age infants may demonstrate lexical knowledge (Mandel & Jusczyk, 1994), which might well supercede any suprasegmental effects. Nevertheless, it is clear that specific f_0 or intensity shapes or contours of utterances carry meaning; a rising and falling contour is quite clearly different in communicative intent from a rising contour. What is not known is the degree to which these contours are specified by either the listener or the speaker. Physiological limits on the respiratory and laryngeal systems, for example, place strict limits on the ranges and rates of change of vocal loudness and intensity. In terms of communication of affective state one might expect that these limits are approached more closely as the speaker's level of arousal increases. Similarly there are well-known limits on the communicative process set by bioacoustic and psychoacoustic limits, which determine auditory sensitivity.

Future efforts will be directed toward extending the present findings to a larger set of speaking conditions and to those exhibiting less exaggerated speech characteristics. Modelling of more subtle f_0 contours will surely present more of a challenge, such that additional models may be considered. Discrimination of utterance types remains a central goal in our investigations, as we see classification of prosodic phenomena as esential to future descriptions and investigations of communication. We further anticipate that these techniques will be exploited for investigations of such phenomena as development of communication in infants as they model their parents' f_0 contours, and in applications of speech signal processing, including synthesis, transmission and compression.

Support for this research was provided by grants from the National Science Foundation (BNS 8919711) and National Institutes of Health (MH 140867, DC00822). The authors gratefully acknowledge the contributions to this work by Lourdes Caro-Martinez, Susan Ranier, Joan West and Adena Zlochower. Portions of these data were presented at the Biennial Meeting of the Society for Research in Child Development (New Orleans, 1993) and at the Annual Meeting of the American Speech-Language-Hearing Association (San Antonio, 1992).

Appendix

The functions to which the f_0 contours were fit are given below. Seven groups are shown, each including one to six functions. Each function within a group yielded a

single regression coefficient, which was averaged with the other coefficients in the classification to yield a single index for the curve type. Note that the power function was used twice; once in a category of its own, and once as a member of the exponential category.

Linear equation

$y=a+bx$ Linear function

Power equation

$y=a+bx^c$ Power function (also included among the exponential equation group)

Decay equation

$y=a+\frac{b}{x}$ Decay function

Peak equations

$y=a+be^{0.5\left(\frac{x-c}{d}\right)^2}$ Gaussian peak

$y=a+\frac{b}{1+\left(\frac{x-c}{d}\right)^2}$ Lorentzian peak

$y=a+be^{\left\{-0.5\ \frac{\ln\left(\frac{x}{c}\right)^2}{d}\right\}}$ Log normal peak

$y=a+be^{\left\{-e^{\left[-\left(\frac{x-c}{d}\right)\right]}-\left(\frac{x-c}{d}\right)+1\right\}}$ Extreme value peak

$y=a+b\frac{4e^{\left[-\left(\frac{x-c}{d}\right)\right]}}{\left\{1+e^{\left[-\left(\frac{x-c}{d}\right)\right]}\right\}^2}$ Logistic peak

$y=a+b\times error\ function\left[\left(\frac{x-c}{d}\right)^2\right]$ Complementary error function peak

Transition equations

$$y = a + \frac{b}{1 + e^{-\left(\frac{x-c}{d}\right)}}$$ Sigmoid transition

$$y = a + 0{\cdot}5b\left[1 + \textit{error function}\left(\frac{x-c}{\sqrt{2d}}\right)\right]$$ Gaussian cumulative transition

$$y = a + \frac{b}{1 + \left(\frac{x}{c}\right)^{d}}$$ Logistic dose response waveform

Waveform equations

$$y = a + b\sin\left(\frac{2\pi x}{d} + c\right)$$ Sine function

$$y = a + b\sin^2\left(\frac{2\pi x}{d} + c\right)$$ Sine^2 function

Exponential equations

$$y = a + be^{\frac{-x}{c}}$$ Exponential function

$$y = a + bx^{c}$$ Power function

$$y = ae^{\frac{-x}{b}} + ce^{\frac{-x}{d}}$$ Two exponential function

References

Bettes, B. (1988). Maternal depression and motherese: Temporal and intonational features. *Child Development* **59**, 1089.

Breznitz, Z. & Sherman, T. (1987). Speech patterning in natural discourse of well and depressed mothers and their young children. *Child Development* **58**, 395–400.

Cooper, W. E. & Sorenson, J. M. (1981). *Fundamental Frequency in Sentence Production*. Springer-Verlag, New York.
Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist* **66**, 103–114.
Fernald, A. (1989). Intonation and communicative intent in mother's speech to infants: Is the melody the message? *Child Development* **60**, 1497–1510.
Fernald, A. & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology* **20**, 104–113.
Frick, R. W. (1985). Communicating emotions through the role of prosodic features. *Psychological Bulletin* **97**, 412–429.
Garnica, O. (1977). Some prosodic and paralinguistic features of speech to young children. In *Talking to Children*. (C. E. Snow & C. A. Ferguson, eds.), Cambridge University Press, Cambridge.
Katz, G. S., Moore, C. A. & Cohn, J. F. (1982). Semi-automated processing of very large natural f_0 samples. Paper presented to the American Speech-Language-Hearing Association, San Antonio.
Mandel, D. R. & Jusczyk, P. W. (1994). Do 4·5-month-olds know their names? *Journal of the Acoustical Society of America* **95**, 3015.
Papousek, M., Papousek, H. & Bornstein, M. (1985). The naturalistic vocal environment of young infants: On the significance of homogeneity and variability in parental speech. In *Social Perception in Infants*. (L. P. Lipsitt, ed.), Ablex, Norwood, New Jersey.
Papousek, M., Papousek, H. & Symmes, D. (1991). The meaning of melodies in motherese in tone and stress languages. *Infant Behavior and Development* **14**, 415–440.
Pierrehumbert, J. (1990). Phonological and phonetic representation. Special issue: Phonetic representation. *Journal of Phonetics* **18**, 375–394.
Price, P. J. (1991). The use of prosody in syntactic disambiguation. *Journal of Acoustical Society of America* **90**, 2956–2970.
Scherer, K. R. (1986). Vocal affect expression: A review and model for future research. *Psychological Bulletin* **99**, 143–165.
Stern, D. N., Spieker, S., Barnett, R. K. & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language* **10**, 1–15.
Stern, D. N., Spieker, S. & MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology* **18**, 727–735.

(Received 14 October 1993 and accepted 27 July 1994)