ELSEVIER

Speech Communication 39 (2003) 367–384

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# BabyEars: A recognition system for affective vocalizations ☆

Malcolm Slaney [a,*,1], Gerald McRoberts [b,1]

[a] IBM Almaden Research, 650 Harry Road, San Jose, CA 95120, USA
[b] Department of Psychology, Lehigh University, Bethlehem, PA 18015, USA

Received 23 January 2001; received in revised form 1 March 2002

**Abstract**

Our goal was to see how much of the affective message we could recover using simple acoustic measures of the speech signal. Using pitch and broad spectral-shape measures, a multidimensional Gaussian mixture-model discriminator classified adult-directed (neutral affect) versus infant-directed speech correctly more than 80% of the time, and classified the affective message of infant-directed speech correctly nearly 70% of the time. We confirmed previous findings that changes in pitch provide an important cue for affective messages. In addition, we found that timbre or cepstral coefficients also provide important information about the affective message. Mothers' speech was significantly easier to classify than fathers' speech, suggesting either clearer distinctions among these messages in mothers' speech to infants, or a difference between fathers and mothers in the acoustic information used to convey these messages. Our research is a step towards machines that sense the "emotional state" of a speaker.

## 1. Vocal expressions of affect

The goal of *affective computing* is to design machines that understand and respond to human emotions (Picard, 1997). Although in its infancy, this field has the potential to alter dramatically the way that humans interact with machines, by allowing machines to adapt their operations in response to users' emotions.

Emotions are complex, transient and ephemeral experiences, comprising both public and private components. Researchers have taken a variety of approaches to the study of affect. Some investigators have looked at the private aspect, measuring the complex internal physiological changes that accompany human emotions; others have examined the public aspect, studying external behaviors such as facial or vocal expressions of emotion. We adopt the latter approach by investigating how various aspects of speech *prosody* (the pitch, rhythm and loudness of speech) are related to vocal expressions of emotion.

Whichever aspect of emotions is chosen for study, researchers must gain access to objectively identifiable emotional episodes that are both genuine and spontaneous. This requirement turns out

☆ This paper presents an extended analysis and description of results that we published earlier (Slaney and McRoberts, 1998). We have reanalyzed our infant-directed data, adding adult-directed utterances, which we consider to have a neutral affective message. We now use standard software for Gaussian mixture-model classifiers, improving performance slightly.

* Corresponding author. Tel.: +1-408-927-1411.

*E-mail addresses:* malcolm@almaden.ibm.com (M. Slaney), gwm3@lehigh.edu (G. McRoberts).

[1] Formerly with Interval Research Corporation.

0167-6393/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(02)00049-3

to pose a significant problem for both logistical and ethical reasons. Either we must follow people around, waiting for pleasant or unpleasant events to occur so that we can capture our subjects' spontaneous responses, or we must manipulate situations so as to elicit such responses. Since both alternatives have obvious practical and ethical limitations, researchers have typically opted to study actors who portray emotional expressions by reading standardized texts in varying tones of voice. Such expressions, however, only approximate genuine emotional expressions. Actors may be good at communicating affect, but their intention is to convey a believable feeling, rather than necessarily to reflect accurately the way that people normally express that feeling. In particular, only a subset of the important cues that people normally use might be present in an actor's portrayal or those present might be exaggerated. Thus, when we study an actor, we know neither what is missing nor what has been added.

A promising way to study genuine expressions of emotions is to observe parents as they speak to their infants; such *infant-directed speech* is often highly affective and is undeniably spontaneous. For example, whether a parent praises a young infant with "Goood girrlll!!" for the baby's first steps, or issues a strong prohibition, "NO! STOP!!" when a toddler is about to pull a lamp off a table, there is little doubt about the affective content, the communicative intent, or the spontaneity of the vocalization. Fernald and her colleagues (Fernald et al., 1989; Fernald, 1989; McRoberts et al., under review) pioneered the study of how adults convey affective and pragmatic messages to infants by recording parents talking to their infants during spontaneous interactions in a variety of naturalistic settings. We adopted this approach to collecting genuine and spontaneous affective vocalizations. Our goal was to see how much of the affective message we could recover using simple acoustic measures of the speech signal.

We studied how adults convey affective messages to infants using prosody. We analyzed speech with low-level acoustic features and discriminated approvals, attentions, and prohibitions from adults speaking to their infants. We built automatic classifiers to create a system, *BabyEars*, that performs the task that comes so naturally to infants. We believe that adult-directed speech can have the same affective messages, and can have the same prosodic patterns, albeit attenuated, as the speech that we studied.

BabyEars discriminates adult-directed (neutral affect) versus infant-directed speech correctly more than 80% of the time, and classified correctly the affective message of infant-directed speech nearly 70% of the time. It judged certain types of speakers' affective messages more accurately than it did others. Specifically, BabyEars classified female speech more accurately than it did male speech. We do not know whether this result was an artifact of our analysis techniques (e.g., our selection of acoustic features) or whether male and female speakers generated different kinds of speech, either in general or specifically in our laboratory environment.

In this paper, we describe previous work on vocal expressions of emotions (Section 2), our data collection (Section 3), our signal-processing techniques (Section 4), and our results (Section 5).

## 2. Affect and prosody

Researchers from psychology, child development, music and engineering have studied the acoustic correlates of vocal expressions of emotion. In early studies (e.g., Skinner, 1935; Fairbanks and Pronovost, 1939) actors spoke prepared texts in varying tones of voice representing different affective states. The results suggested that various prosodic parameters, such as pitch range, intensity and speech rate, correlate with the different affective tones. Later studies using similar approaches (e.g., Williams and Stevens, 1972; Scherer, 1986) confirmed a role for prosodic parameters. However, the informativeness of these studies was limited by the use of univariate statistical measures and by methodologies that relied on the use of prepared texts rather than spontaneous speech.

Recently, more sophisticated acoustic analyses and multivariate statistical classifiers have been applied to the study of vocal expressions of emotion. For example, Roy (Roy and Pentland, 1996) analyzed the pitch, energy and spectral tilt of recorded positive and negative messages; linear discriminant analysis correctly classified the affective valence 70–75% of the time.

We used infant-directed speech in our study because of its clear and unambiguous style. Researchers studying infant development and language acquisition observed that, when mothers in various cultures speak to infants, they often use a special speech register that includes higher pitch and wider pitch range (e.g. Garnica, 1977; Ferguson, 1964). Fernald highlighted the significance of this "infant-directed" speech prosody in several studies. One study demonstrated the ubiquitousness of infant-directed speech prosody by showing that parents (mothers and fathers) from various cultural and language backgrounds consistently used higher pitch and wider pitch range when talking to their infants than when talking to adults (Fernald et al., 1989). Another study focused on the communicative aspects of infant-directed speech prosody, showing that infants attend longer and smile more when listening to the exaggerated intonation of infant-directed speech than to adult-directed speech by the same speaker (Fernald, 1985).

These findings and others are consistent with the hypothesis that there are two important communicative functions of infant-directed speech prosody. One is an attentional function: The prosody of infant-directed speech serves to engage and maintain the attention of infants. The other is the communication of affect or emotion. Since infants smile more when listening to infant-directed speech, it appears that parents are often communicating positive affect in their speech to their young infants.

In another study, Fernald (1989) suggested that the prosodic cues in adults' affective expressions to infants are clearer than are cues in similar expressions to adults. In Fernald's study, adults were recorded speaking spontaneously to their infants and then to their spouses in role-playing situations calling for similar affective messages (approval, prohibition, comfort, attention bid and game initiation). After the researchers low-pass filtered the utterances at 400 Hz to remove most of the formant information, listeners were able to identify the affective message in 60–80% of the infant-directed utterances, but in only 30–40% of the adult-directed utterances. Therefore, it appears the affective messages in infant-directed speech are expressed more clearly or understandably than in adult-directed speech. In a related study, Fernald showed that infants respond with appropriate affect to infant-directed approvals and prohibitions, even when the language is unfamiliar to them (Fernald, 1993). Thus, these messages may be expressed consistently across languages and cultures.

Although these studies show that some affective messages are clearly expressed in infant-directed speech and are responded to by young infants, they leave open the question of precisely how these messages are communicated. One approach to answering this question is to study the shape of the intonation contour of utterances that express various communicative functions, such as praises and prohibitions. Several studies have used this approach, typically by recording spontaneous infant-directed speech and then classifying the pitch contours of the various utterance types into categories such as rising, falling and bell shaped (Papousek and Papousek, 1991; Stern et al., 1982). The results of these studies are difficult to interpret, in part because of the subjectiveness of establishing the contour shape, but they do suggest that specific intonation contour shapes may be associated with specific communicative intents. For example, Stern and Papousek report that bell-shaped or rise–fall contours often are used by a parent who is praising the infant. Fernald (1992) concurs, and also suggests that prohibitions are often associated with short, falling contours. However, the predictive power of these observations is unclear, because these studies did not include any way to compare the categorization or classification power of the contour shapes.

A different approach to relating prosodic form to communicative function in infant-directed speech is to look at a variety of prosodic features, such as mean pitch (F0) and pitch range, rather than to focus on intonation contour shapes. Taking a cue from studies of animal communication, McRoberts, Fernald and Moses (McRoberts et al., under review) used discriminant analysis to establish the ability of 24 prosodic features to distinguish among three types of infant-directed utterances (attention, approval, prohibition) made by mothers in four language groups. Not only was this approach successful within individual languages, but also a single model incorporating just eight prosodic features categorized 58–69% of utterances successfully across all four languages.

Katz and his colleagues (Katz et al., 1996) compared the ability of prosodic features and intonation contours to classify three communicative functions in infant-directed speech (attention, approval and comfort). They used curve-fitting techniques to remove some of the subjectiveness of classifying intonation contours in previous studies. Multivariate statistical analysis correctly classified 69% of the utterances. The contours and features had equal predictive power in these classification tests.

Note that infants operate on aspects of the vocal signal different from those used by either adults or current speech-recognition systems. Infants understand the prosodic message long before they understand the linguistic message, whereas speech-recognition systems generally analyze the words and ignore the prosody (Fernald, 1993). For adults, the words and prosody of an utterance normally contribute to both the linguistic and the affective message. For example, the words "yeah, right" can convey a positive message or a resigned negative message, depending on pitch and timing. In many situations, minimal prosodic information conveys much information about the intent of the message. Thus, to improve the recognition and comprehension of speech by computers, some speech researchers have added information about the prosody of speech to conventional speech-recognition systems information about the prosody of the signal (Price et al., 1991).

Our work builds on previous work in several ways. We use infant-directed speech as an example of a universal emotional communication. As we describe in Section 3, our data are from spontaneous and natural speech that is representative of universal affective communication. To classify the emotional content of this speech, we built the statistical classifiers described in Section 4. Because the prosody that adults use when speaking to infants appears simple, some researchers have suggested detecting and classifying the shape of the pitch contour. Thus a sharp upward rise in pitch might mean something different from a long, slow glide. BabyEars does not do this kind of analysis directly. However, our technique of splitting the utterance into thirds allows us to capture some of this information and then a statistical classifier can determine the utility of the information.

## 3. Data collection

Our study comprised two experiments. First, we collected acoustic data from parents talking to their infants. Then, different adult listeners judged (1) whether each utterance was best classified as an approval, attention or prohibition, and (2) the strength of the message. We initially considered only infant-directed speech, but later reanalyzed our recordings to extract neutral, or adult-directed utterances.

### *3.1. Acoustic data*

We recorded six mothers and six fathers while each was interacting with their 10 to 18-month-old infant in a quiet room. Each recording session lasted about 1 h, during which each parent was asked to play and otherwise interact normally with their child. Several toys were placed in the room. We asked the parents to use only their voices to keep their child away from several "dangerous" items, such as lamps and microphones. An experimenter stayed in the room to oversee the experiment and to encourage verbal interaction. The audio samples that we collected were spontaneous and natural.

We recorded two channels of audio. A lightweight headset microphone (Countryman hypercardioid Isomax headset microphone, and a Sony WRT-820 radio transmitter) was used as the primary source for the parents' voices. A second microphone recorded the audio from all the participants (experimenter, child, and parent). Both channels were recorded directly to a computer's hard disk at 16-bit resolution at a 44 kHz-sampling rate, and then were downsampled to 22 kHz.

The results reported here are for only those data collected with the headset audio channel. (The second, environmental microphone channel was reserved for future research.)

A trained research assistant extracted discrete utterances, placed each into its own disk file, and classified each as being in one of three classes of vocalizations: approval, attention and prohibition. (We originally included a class of data for encouragement, but dropped it when we found that we could not select appropriate utterances consistently.) Utterances were defined as a single sentence or phrase bounded by silence longer than the interword gaps and were from 0.53 to 8.9 s long. Although adult listeners independently classified each utterance (see Section 4.2), only the segmenter had access to the surrounding audio; she thus was assumed to be able to classify the meaning of each utterance more accurately than the listeners in our second experiment. Therefore, we used the segmenter's labels as true values.

Typical examples of the three categories that we use follow (the prosodic contours are, of course, missing):

Approval: "Wow!" "Yea. Good Boy."

Attention: "Becca!" "Nicholas, here!" "Anthony?"

Prohibition: "That's not for you." "Don't go in there!"

For each parent–infant pair, we selected and classified 30–50 utterances, for a total of 509 infant-directed utterances (212 approvals, 149 attentions and 148 prohibitions).

We selected the infant-directed speech samples based on the criterion that they have a clear affective message. For many applications, however, we want to know both *whether* there is an affective message as well as *what* that message is. To answer the first question, we needed samples of speech that had little or neutral affective content. Thus, we selected an additional 185 adult-directed utterances from the original audio files. Most such utterances were simple informational statements or questions directed to the experimenter in the test room. We did not pick any utterances where the parent was mimicking the infant. We analyzed a total of 694 utterances.

### 3.2. Adult listeners' assessments

We asked seven adults, unrelated to the parents in the initial experiment, to listen to each utterance, both to measure their agreement with our segmenter's classifications and to get an estimate of the strength of the affective message. These adults, who had no training in either linguistics or psychology, listened to each segmented infant-directed utterance and judged its category (approval, attention or prohibition) and strength (on an arbitrary scale from 1 (weak) to 5 (strong)). The listeners did not know our research hypothesis and were neither trained nor instructed to listen specifically to prosody.

## 4. Analysis

BabyEars analyzes speech using three classes of features: pitch, broad spectral shapes and energy variations. In brief, we hypothesized that speech that was slowly and smoothly varying would indicate approval, whereas sounds that changed quickly would be attention bids or prohibitions. BabyEars measures several variations of these parameters. It uses multi-dimensional Gaussian mixture models (GMMs), described in Section 4.2, to model the probabilities of the data, and to choose the most likely affective class. We analyzed BabyEars' ability to classify the 694 utterances.

### 4.1. Signal processing

BabyEars processes each manually segmented sound file automatically at a frame rate of 50 Hz. A speech–silence discriminator further segments each sound file at phrase boundaries (Lamel et al., 1981). BabyEars chooses the longest phrase in each file for additional processing. We evaluated our performance using the speech–silence discriminator, instead of relying on human segmentation, because we wanted a completely automatic system for classifying affect from an open microphone.

For analysis, BabyEars processes each utterance as a whole, then splits each utterance into thirds (beginning, middle, and end) by time duration as shown in Fig. 1. Thus, for each feature—for

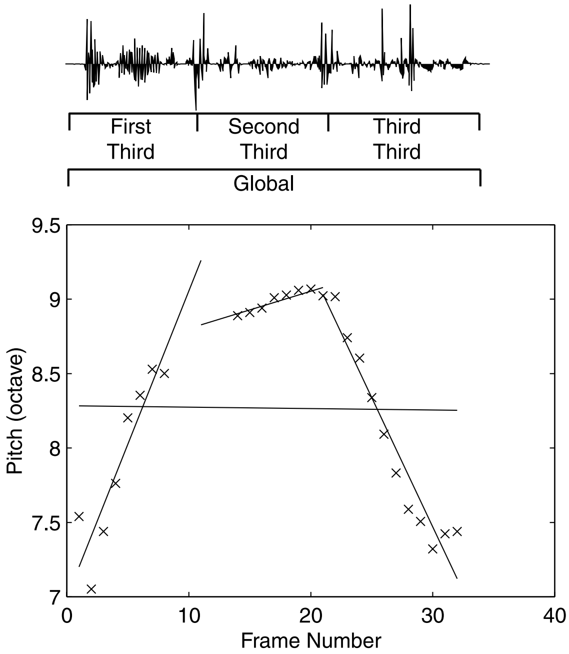

Fig. 1. We computed features based on three different portions of each utterance and the entire (global) properties of the utterance. The lower graph shows the pitch estimate for one utterance, with the four slope measurements shown.

example, the pitch range—we have four measurements over different time periods. BabyEars computes three types of analyses on each temporal period of each utterance: pitch, cepstral or spectral-shape changes, and energy. Detailed descriptions of the eight analysis measures that BabyEars used in this study are given in Appendix A.

BabyEars analyzes the pitch of each utterance using a high-quality dynamic-programming algorithm (Talkin, 1995). The pitch module produces estimates of the speech signal's pitch, measured in Hertz. BabyEars then computes the base 2 log of this number to collapse the pitch estimate into octaves and to put the measurement on a perceptual scale. BabyEars does not do any postprocessing to correct for possible octave errors. We chose Talkin's pitch detector because it gave the fewest octave errors in our informal tests. [2] No pitch detector is perfect. The results presented here should be considered a lower-bound on performance: Fewer pitch errors will make the classifier's job easier and the performance higher.

BabyEars computes several statistics related to the pitch: the variance, slope, range (maximum minus minimum) and mean. BabyEars also measures two statistics of the frame-by-frame delta pitch: the mean delta pitch, and the mean of the absolute delta pitch. The mean delta pitch is similar to the slope measurement. When either frame's pitch is undefined, because it is unvoiced, the delta-pitch measures are undefined and do not enter into the calculation.

BabyEars uses mel-frequency cepstral coefficients (MFCC) (Hunt et al., 1980) to characterize the broad spectral shapes in the utterance. MFCC parameters are often used in speech recognition as a simple representation of the acoustic waveform. We use MFCC as an acoustic measure of the articulatory configuration. We wanted to investigate whether the speed with which these parameters changed would be a useful feature for differentiating between prohibitions and approvals. Thus, BabyEars measures the mean frame-by-frame change in the MFCC parameters during each segment of the utterance. In this calculation, BabyEars ignores the energy, or C0 component, and sums the absolute value of the changes in the remaining coefficients.

Note that MFCC does not attempt to measure the formant frequencies directly. Instead, MFCC gives a multi-dimensional representation of the vocal-tract configuration, much like linear-predictive coding does. We use MFCC in this work because it has allowed researchers to succeed in previous speech-recognition experiments.

Finally, BabyEars also computes the variance of the energy in dB in each frame, across each utterance.

### 4.2. Classification

BabyEars uses a multidimensional discriminator to assign each utterance to a class. We judged BabyEars' performance on its ability to assign the same affective label that humans used.

[2] Droppo (Droppo and Acero, 1998) reports that the standard deviation of errors from Talkin's pitch detector was from 0.34–0.74% in their tests.

We trained the GMMs that BabyEars uses to recognize our data. In all results reported in this paper, we used the Netlab GMM software (Nabney, 2001) in MATLAB. In a GMM, we find a set of multidimensional Gaussians that combine to model the overall probability of one class of data in the given feature space. Thus, for each class of data, we form an estimate of the probability that the feature vector assumes the value $x$ (in a $d$-dimensional space) by adding $M$ of these Gaussian "bumps" (Bishop, 1995),

$$P(x) = \sum_{j=1}^{M} P_j \phi_j(x).$$

Here $P_j$ is the weighting of the $j$th Gaussian and $\phi_j(x)$ is a Gaussian probability given by

$$\phi_j(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \times \exp\left[-\frac{1}{2}(x-\mu_j)^t \Sigma_j^{-1}(x-\mu_j)\right],$$

and $\mu_j$ and $\Sigma_j$ describe the Gaussians that we used to model this class of data: $\mu_j$ is the mean, and $\Sigma_j$ is the covariance. Each Gaussian has a diagonal covariance, filling the prosodic vector space with Gaussian ellipses, each with its axes aligned with the coordinate system. We build a classifier by comparing the posterior probabilities from each model and choosing the largest. (This model assumes that all classes are equally likely to occur and that all mistakes incur the same cost.)

The modeling power of a GMM is related to the number of Gaussians, $M$, that we use to model each class of data. More Gaussians produce more accurate models of the data, but also carry an increasing risk of overfitting the training data. BabyEars' performance as a function of the number of Gaussians, is shown in Fig. 2. Based on this result, we use 10 Gaussians per model in all the remaining trials because classification performance was near the maximum. We obtained similar results with optimal linear discriminators.

We built optimal classifiers using greedy selection. At each step, we trained three sets of GMMs, one set for each class, with the current set of features and each remaining feature. We then chose the feature that resulted in the best performance, and added that feature to the set. In this way, we found an approximation to the best features for making this classification. This test gave us information about which features were adding the most information to the classification decision.

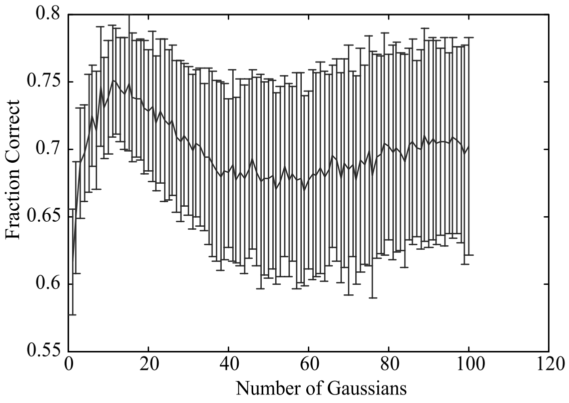

Fig. 2. Classification performance in a three-way test as a function of the number of Gaussians used to model each class' probability distribution. When we used a small number of Gaussians, we did not have the power to capture the true distribution; a large number of Gaussians resulted in overtraining. Both extremes lowered performance. We have plotted one standard-deviation error bars calculated by replicating our training and testing procedure 100 times in the "0.632" paradigm using all 32 features and our very strongest affective data (see Fig. 4).

The errors that we report are naturally biased by the relative frequency of the affective classes in our data set. Since test samples were drawn at random from the data that we collected, we tested with more approval utterances than prohibition utterances. Although this choice slightly biased our results, we do not believe that it changes our overall conclusions.

### 4.3. Bootstrapping

Since we had a limited set of data, we used the "0.632" bootstrapping procedure (Efron and Tibshirani, 1993) to estimate BabyEars' performance and the variance of our estimates. We trained our classifiers with a set of data, chosen randomly with replacement, that was equal in size to the original data set. We tested BabyEars' classifiers with all data that were not used in training. We repeated this task 100 times per discriminator, and then averaged the results to find an estimate of the mean and standard deviation of the recognizer's performance with that set of features.

In bootstrapping, we estimate the true performance of a classifier as though we had used all $n$ training samples to classify an utterance that we had not seen yet. Training and testing on all the data would give an overly optimistic result. Jackknifing procedures leave one datum out of the training set and then tests performance on that datum. Jackknifing gives accurate estimates of the performance, but with large variances, because every test result is all correct or all wrong. Bootstrapping uses a portion of the data for training and the remaining data for testing. By extrapolating our results, we get an estimate of the true classifier performance that we would get if we used all the data to build the classifier, and we also get an estimate of the variance of our performance estimate.

In the 0.632 bootstrap procedure, we first select $n$ training points, with replacement, from the original $n$ data points. Because we are choosing with replacement, the probability distribution of the training data is an unbiased approximation of the original data's probability distribution. It is easy to show that for large databases, on average $1 - 1/e = 0.632$ of the original data are chosen for training; some data points are chosen multiple times. By repeatedly choosing bootstrap samples and training classifiers, we can obtain reliable estimates of the true error rate and can also get a confidence estimate.

There are many ways to estimate the error of a classifier. We compute the apparent error, $e_{\mathrm{a}}$, using the same data for training and testing; the apparent error is an optimistic estimate of performance, since the classifier is tuned to the testing data. The error from one bootstrap sample, $e_{\mathrm{b}_i}$, is pessimistic; the classifier is trained on a subset of all the available data, and we would certainly do better if we used all the data. The true error rate—$e_{\mathrm{t}}$, the errors we would see if we used all the data to train our classifier and then tested on new data—is a value between these two estimates. The 0.632 bootstrap procedure estimates the true error of the classifier as

$$e_{\mathrm{t}} = 0.368e_{\mathrm{a}} + 0.632\frac{1}{k}\sum_{i=1}^{N} e_{\mathrm{b}_i},$$

where each of the $N$ bootstrap trials gives a classification error of $e_{\mathrm{b}_i}$. There were $N = 100$ bootstrapping trials in all the studies reported in this paper.

We use the distribution of bootstrap-error estimates to give an estimate of our confidence in the classifier's performance. The lengths of the error bars in all figures indicate plus and minus 1 standard deviation of the distribution of the errors that we see during the bootstrap procedure. Since the bootstrap training sets are not independent, we underestimate the true variance.

## 5. Results

We tested the performance of our adult listeners and BabyEars' classifiers in several different ways. Section 5.1 describes how adult listeners classified the test utterances. Section 5.2 reports BabyEars' multivariate classification performance for several models of the infant-directed speech. Section 5.3 presents the confusion data. Section 5.4 shows simplified decision surfaces for our data, which provide an understanding of the acoustic parameters that BabyEars uses to make decisions. Section 5.5 describes how well affective recognizers trained on male or female speech generalize to the other gender. Section 5.6 describes BabyEars' performance when the adult-directed or neutral speech is added to the recognition task. Section 5.7 describes the correlation of BabyEars' performance in recognizing adult- and infant-directed speech. Finally, Section 5.8 discusses the performance of speaker-dependent classifiers.

### *5.1. Listener ratings of infant-directed speech*

Our listeners were not 100% consistent in classifying the utterances; All seven agreed unanimously with our segmenter's classifications for 79% of the infant-directed utterances, at least five of the seven listeners agreed with the segmenter for 85% of the utterances and at least four of seven listeners agreed with segmenter's classification for all utterances.

The listeners also varied in their judgments of the strength of each utterance. Their overall mean strength rating of the infant-directed utterances was 3.2; approval utterances received an average strength rating of 3.6; attentions and prohibitions were scored 2.8 and 2.9, respectively. Perhaps our parents were more willing to issue strong approval messages than they were to get their child's attention (and startle her) or to prohibit her. Utterances judged to be weak in affective content often contained a linguistic message that did not match the prosodic message. For example, "Nicholas, don't do that" said with a soft, pleading voice is a linguistic prohibition but with an encouraging (or perhaps resigned) affective message.

In addition to the entire data set, we analyzed two subsets of data based on the adult listeners' classifications and on these listener's perceptions of affective strength. Both subsets included only those utterances whose affective message was considered clear and strong, defined as agreement by at least five of the seven listeners with our original classification. Additional criteria were based on the average strength rating of the adult listeners; we hypothesized that utterances with strong affective messages would be easier to recognize.

Thus, the three sets of infant-directed utterances were as follows:

- *All Data:* This set included all utterances, including those for which the listeners did not agree with our original classifications. Based on our original classifications, there were 212 approvals, 149 attentions and 148 prohibitions. There were 263 utterances from female speakers and 246 utterances from male speakers. There were 509 utterances in this class.
- *Strong Data:* This set included utterances for which at least five of the seven listeners agreed with our initial classification and that had an average strength rating greater than 2.5. There were 208 approvals, 110 attentions and 112 prohibitions that were assigned this strength; 220 were from female speakers, and 210 were from male speakers. There were 430 utterances in this class.
- *Very Strong Data:* This set included utterances for which at least five of the seven listeners agreed with our initial classification and that had average strength ratings greater than 3.0. There were 179 approvals, 64 attentions and 75 prohibitions; 165 were female, and 153 were male. There were 318 utterances in this class.

For the three categories, the adult listeners agreed with our segmenter's classifications for 96% (All Data), 97% (Strong Data) and 98% (Very Strong Data) of the utterances. (These percentages are higher than those described at the start of Section 5.1 because they measure average error rate, not uniformity of judgement.)

### 5.2. Multivariate classification

Fig. 3 shows the BabyEar's classification results for the individual acoustic features of the infant-directed speech. Classification using any one of many of the features gave performance greater than chance. Three features stand out: Two global pitch measures and the delta MFCC feature each produced about 50% correct classification—significantly better than chance in this three-way test. These results, and those that follow, were all computed using 10-Gaussian mixture models to

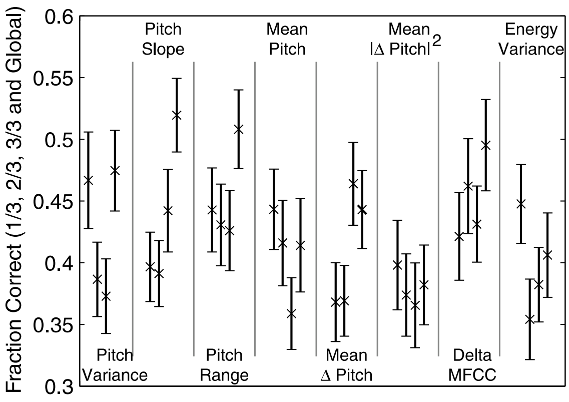

Fig. 3. Univariate classification performance in a three-way test for each of the 32 features that we studied. For each feature type, we show the percent correct for four different temporal regions of each utterance—first, middle and last third; and the entire utterance—as shown in Fig. 1.

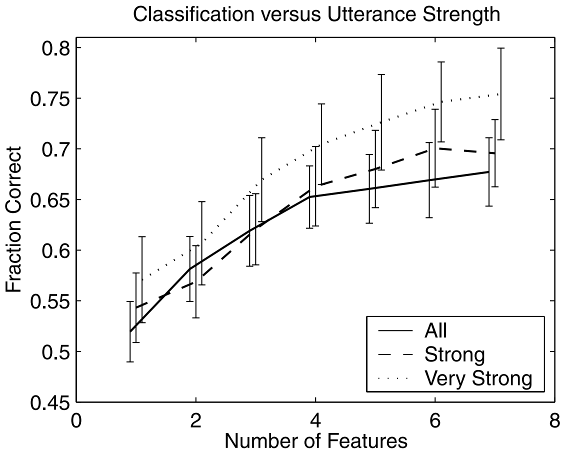

Fig. 4. Classification performance in a three-way test (approval, attention, prohibition) as a function of the number of features. Three curves show our performance for three sets of our data. Utterances with a strong and clear affective message (dotted line, five of seven adult listeners agreed on the classification) were easier for BabyEars to recognize.

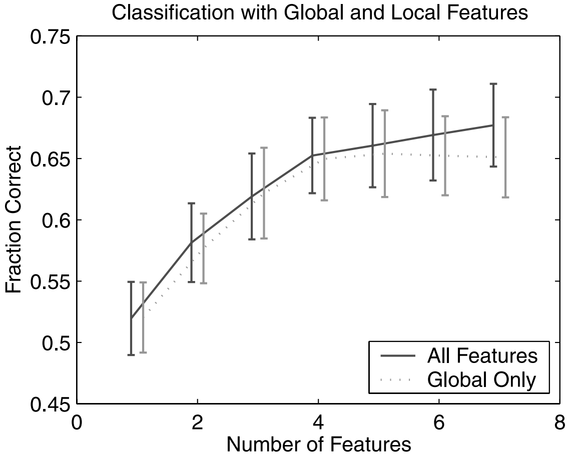

Fig. 5. Prosodic features were calculated over thirds of the utterance and over the entire utterance (see Fig. 1). This figure shows the performance of a three-way classifier (approval, attention, prohibition) at distinguishing these messages using all features and just the global features. The local (third-utterance) features did not perform significantly better than the global features.

represent the probability distribution of each class of data.

We found the optimal classifier by starting with the single best feature and adding features one at a time, at each step adding the feature of those remaining that provided the greatest improvement in performance. Fig. 4 shows classification results for all our infant-directed training data as we added more features to the classifier. Classification performance increased as more features were added, then leveled off above 67% with five to seven features.

Classification performance improved when we considered only those utterances that were judged by our listeners to have strong and unambiguous affective messages. Fig. 4 compares classification performance when bootstrapping included All Data, Strong Data and Very Strong Data. Classification results were highest when the data set was limited to vocalizations with the highest average strength ratings (i.e., Very Strong Data). Appendix B lists the features used for all classifiers.

The results shown in Figs. 2 and 3 are based on a human segmenting discrete utterances from the speech stream and then further segmenting individual utterances into thirds (see Fig. 1), an approach originated by McRoberts (McRoberts et al., under review). Unfortunately, segmenting speech into discrete utterances so that they can be split accurately into thirds is difficult. We can avoid this problem by using only global features, which are less sensitive to how well the speech is segmented, in the classifiers. Fig. 5 compares models using all features with those using only global features. Recognition rates were slightly—but not significantly—lower with global features, corroborating the hypothesis that prosodic shape is not important. (Note the "all" curve in Fig. 5 is identical to the "all" curve in Fig. 4.)

Fig. 3 shows that both pitch and articulatory information are useful to classify an affective message. As shown in Appendix B for the strongest utterances in our database, the top two features chosen are MFCC and pitch slope, indicating these two measures are providing relatively independent information about the speaker's message.

### 5.3. Classification confusions

Classification performance varied with affective message. Overall performance is shown in Fig. 6 for the three types of infant-directed messages.

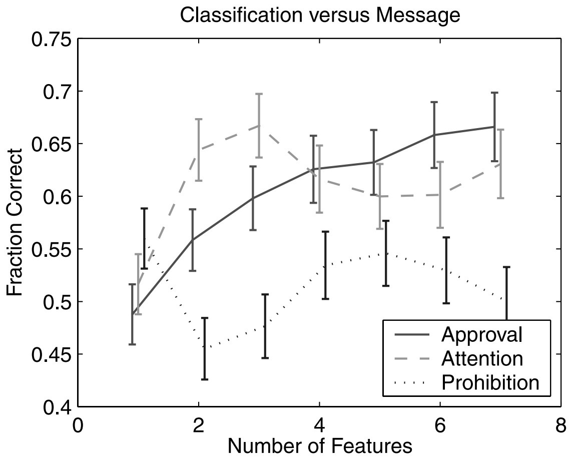

Fig. 6. Classification results for three different types of affective messages. Prohibitions were classified correctly significantly less often than were either approvals or attentions.

Table 1
The confusion matrix for three-way classification tests on All Data based on 20 experimental runs

| Truth\ measured | Approval | Attention | Prohibition | Correct |
|---|---|---|---|---|
| Approval | 1043 | 176 | 317 | 67% |
| Attention | 173 | 685 | 206 | 64% |
| Prohibition | 306 | 197 | 563 | 52% |

Table 2
The confusion matrix for three-way classification tests that used only Very Strong Data based on 20 experimental runs

| Truth\ measured | Approval | Attention | Prohibition | Correct |
|---|---|---|---|---|
| Approval | 1058 | 98 | 154 | 80% |
| Attention | 145 | 271 | 63 | 56% |
| Prohibition | 245 | 68 | 241 | 43% |

Tables 1 and 2 compare the confusion matrices for a classifier built with All Data and for one built with only the Very Strong Data. Each classifier was built with the seven best features shown in Fig. 4. In both cases, BabyEars recognized approvals and attentions moderately well, but its performance was noticeably weaker for prohibitions.

The confusion-data results in this paper (Tables 1–3) represent the raw errors used as input to the bootstrapping procedure. As such, they have an error rate higher than that shown in Fig. 4: A portion of the errors arises because each classifier was built with only 63.2% of our training data.

Table 3
The confusion matrix for four-way classification tests on only Very Strong Data based on 10 experimental trials on the testing data

| Truth\ measured | Approval | Attention | Prohibition | Adult | Correct |
|---|---|---|---|---|---|
| Approval | 427 | 72 | 75 | 120 | 61% |
| Attention | 48 | 142 | 52 | 12 | 55% |
| Prohibition | 97 | 55 | 65 | 60 | 23% |
| Adult (neutral) | 188 | 19 | 54 | 372 | 58% |

The confusion data in Table 1 (All Data) show that prohibitions are mistaken for approvals and attention bids. As shown in Table 2 (Very Strong Data), prohibitions were primarily confused with approvals. One possible explanation for the poor performance in classifying prohibitions is that parents were unwilling to give extremely strong prohibitions in our laboratory while being recorded; whether because they were afraid that the experimenter would not approve, or because they did not want to upset their children, or because no situation warranted it. In Fernald's study (Fernald, 1989), prohibitions were also difficult for adults to recognize, even though the recordings were made in the infant's home.

Our classifiers' performance for approvals increased as it looked at stronger utterances, indicating that these utterances were more distinct from the other classes and easier to discriminate. In contrast, its recognition rate for attention and prohibitions decreased with utterance strength, perhaps indicating that more of the affective message lay in the semantics than in the prosody.

The drop in performance for prohibitions could indicate mismatches between the linguistic and prosodic messages not differentiated by our adult listeners, resulting in a higher proportion of strong linguistic and weak prosodic prohibitions in the Strong and Very Strong Data than in All Data. Perhaps adult's approval utterances do not contain such mismatches.

### *5.4. Decision surfaces*

A classifier looks at the input data—up to 32 acoustic features per utterance in this study—and makes a decision. Whatever technology it uses to make that decision, it chooses a definite winner for each point in the *d*-dimensional space. The boundary between one class of data and another is a *decision surface*. In an optimal Fisher discriminator, the decision surfaces are hyperplanes; in contrast, the GMM classifiers in this study split the space along elements of hyperellipsoids.

BabyEars models the distribution of each class of data (Fig. 7) with a mixture of Gaussians and the relative probability of each mixture determines the decision surfaces (Fig. 8). Each of the strongest female utterances is plotted as a function of four commonly used global features: pitch range, pitch slope, delta MFCC, and energy variance. We used the infant-directed data for the upper two plots, and then added the adult-directed data in the lower plots.

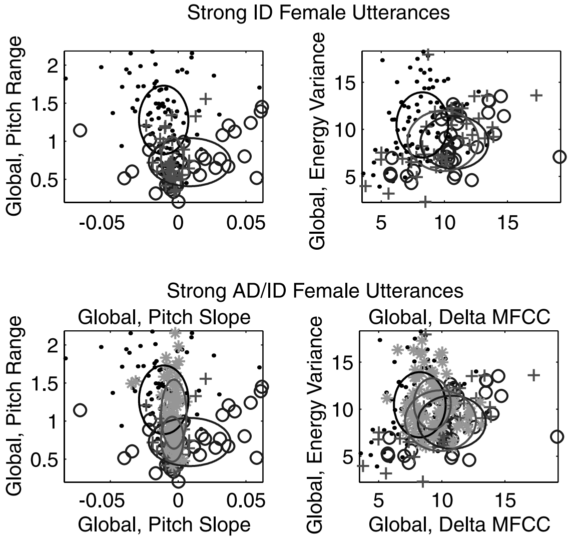

Fig. 7. Each of our strongest utterances plotted as a function of two pitch variables (left) and energy and MFCC (right). The top two plots show the measurements for infant-directed speech; the bottom two plots also include the measurements of the adult-directed, or neutral, speech. Approvals are dots, attentions are circles, prohibitions are crosses, and adult-directed or neutral are stars. The ellipses represent the one standard deviation boundary of each class of data.

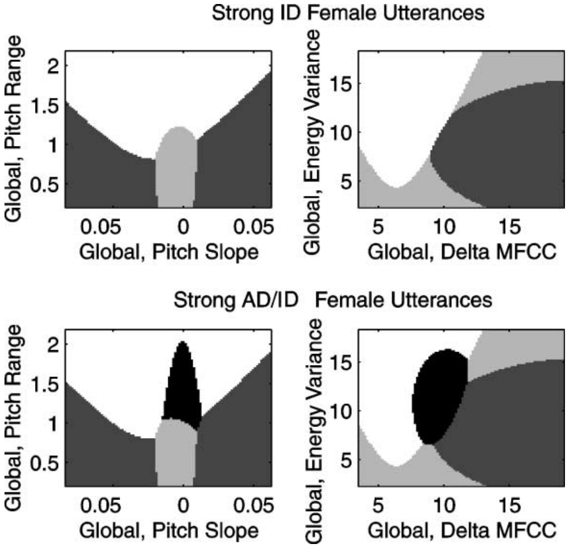

Fig. 8. Decision surfaces for the strongest utterances of each affective class based on single Gaussian models. The graphs show the regions on a two-dimensional plane for which the probability of any given class is larger than the probability of any of the others. See the caption of Fig. 7 for details. Regions of these plots which are white are judged to be approvals, light gray regions are prohibitions, dark gray regions are attentions and black regions are adult-directed utterances.

Fig. 7 shows the raw data as a scatter plot. Many of the data overlap in the center of these plots; there are regions where one type of utterance is much more prevalent than are the others. Decision surfaces from large numbers of Gaussians are difficult to interpret. For this reason, although we used 10 Gaussians to calculate the classification results given elsewhere in this paper, in Fig. 7 we model the data for each class with a single Gaussian and plot the one standard-deviation contour.

Fig. 8 shows most likely affective message for possible values of these four variables. These decisions were based on the single-Gaussian model that best represents the raw data shown in Fig. 7: Each pixel of the plot is colored based on the Gaussian model which had the highest probability for the values of the feature represented by the pixel's location.

Many features help us to build good classifiers. Figs. 7 and 8 show the results when BabyEars used the global pitch range and slope, delta MFCC, and global energy variance: Classifiers built with these features recognized 61%, 62% and 67% of utterances in the three sets of data shown in Fig. 4. The optimal classifiers, which chose from all 32 features, correctly recognized 65%, 66% and 70% using the four optimal features.

Fig. 8 shows the decision surfaces that resulted when BabyEars used a single Gaussian to model each of the three classes of data. At each point in the vector space, we evaluated the probability of each model, and then colored the appropriate pixel in the image corresponding to the winning class. This simplified view of our data shows several trends.

First, prohibitions have little pitch slope; indeed, most of the prohibitions have a flat pitch. Approvals and attentions are disambiguated by separate pitch cues. Attentions have a large pitch slope, which can be either positive or negative. On the other hand, approval messages have a large pitch range.

Second, adult-directed speech tends to lie at the intersection of the three infant-directed classes. This finding is consistent with the idea that our adult-directed speech has a neutral affect compared to the strongest infant-directed messages.

### *5.5. Male versus female features*

For reasons that we do not understand, affective classification performance was higher for female speakers than for male speakers (Fig. 9). The female utterances were classified at a rate up to 76% correct, whereas male utterances were classified correctly at best 64% of the time. This difference could be caused by any of five factors.

1. The acoustic features that we analyzed may not be optimal for male speech.
2. Our procedure may not capture the full affective range of the male speakers.
3. Female speakers may be more skillful or more practiced in producing characteristic infant-directed speech.
4. Male speakers may have been less willing or able in our corporate laboratory environment to produce the prototypical utterances of infant-directed speech.

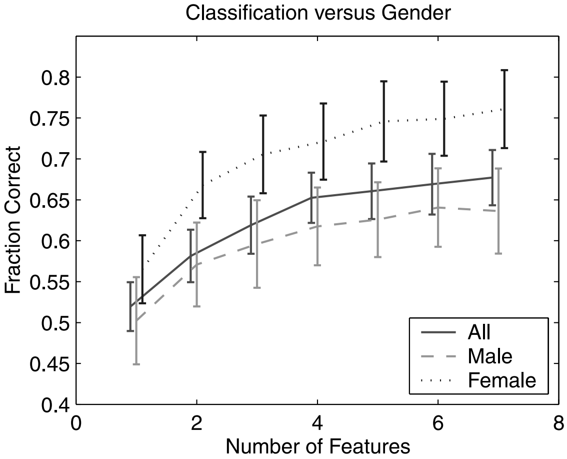

Fig. 9. Gender-specific classification performance in a three-way test (approval, attention, prohibition) as a function of the number of features.

5. People and BabyEars both may perceive male speech differently from the way they do female speech.

Interestingly, average strength ratings by our listeners were equal for male (3.14) and female speakers (3.15). Perhaps male speakers were putting more of the message into the words than into the prosody, so the average strength of the utterance was judged the same as female speakers who did the opposite.

Male and female speakers encoded their infant-directed speech in different ways; the features and classifiers optimized for one gender worked poorly on the other gender. Fig. 10 shows the results of several tests comparing training and testing on the two genders. In the first case (infant-directed speech), models trained on male utterances performed poorly but nearly identically on utterances by both genders. In most cases, the features and discriminators trained with utterances made by one gender generalized poorly to utterances made by the other gender. In the two-way tests of adult-directed versus infant-directed speech, classifiers optimized for utterances made by one gender's classifier produced results less accurate than chance when applied to utterances by the other gender. These results (and those shown in Fig. 9) indicate that gender-independent recognition is

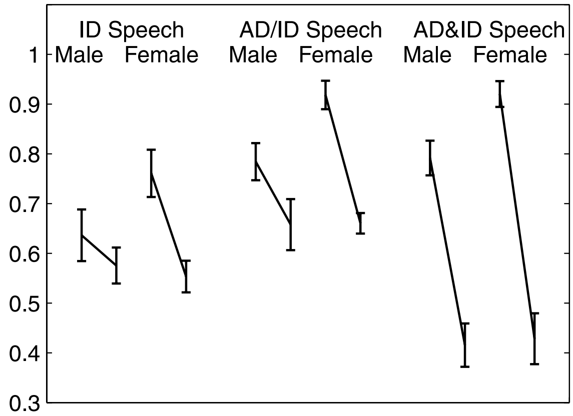

Fig. 10. Comparison of gender-specific models for three different kinds of classification experiments. Each line shows how performance falls when a model optimized for one gender is applied to utterances made by the other gender. Each line is labeled with the gender of the speaker of the training data. The first pair of lines uses the four best features in Fig. 9, the second pair uses the features from Fig. 11, and the final pair uses the features from Fig. 12.

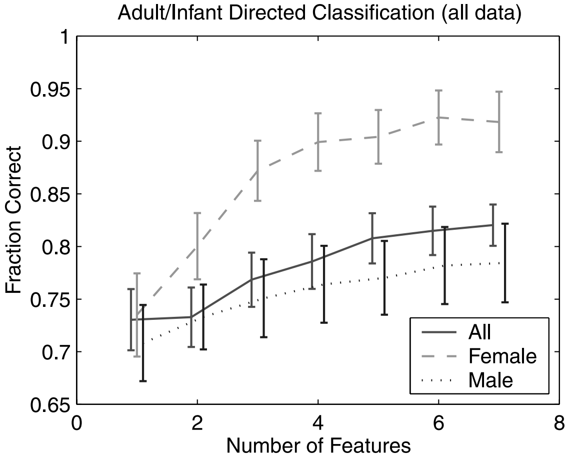

Fig. 11. Adult-directed (neutral) versus infant-directed (affective-message) classification performance as a function of the number of features.

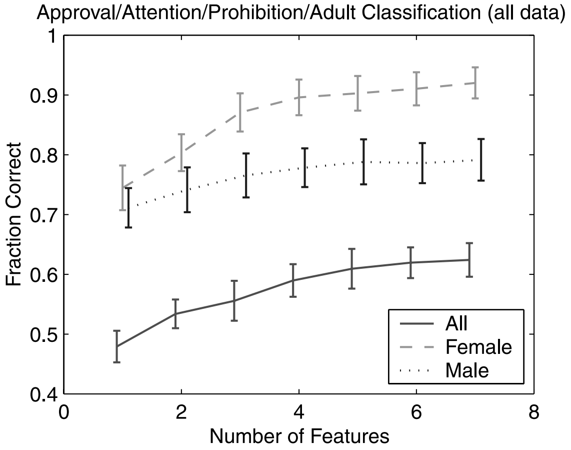

Fig. 12. Four-way (approval, attention, prohibition and neutral) classification results as a function of the number of features.

possible when the system is trained with both male and female examples; perhaps there is enough information in the pitch range of each utterance to disambiguate the gender. More importantly, these results suggest that the male and female speakers in our study are conveying affect with different prosodic features.

### 5.6. Adult-directed versus infant-directed speech

We studied adult-directed utterances in separate tests. Fig. 11 shows BabyEars' performance discriminating between adult-directed and infant-directed speech, or equivalently detecting speech that had an affective message. In this test, we used all the selected adult-directed and infant-directed speech. Again, we were able to classify the female speech more accurately (92%) than the male (79%) in this two-way test.

Results for four-way classification tests—neutral, attention, approval and prohibition—are shown in Fig. 12. In this case, both male and female utterances were classified more accurately when they were separated by speaker's gender than when all utterances were combined. The inability of one gender's recognizer to classify the other gender's utterances is strong evidence that our classifiers were using different features to make the distinctions, possibly because male and female speakers were communicating their affective messages in different ways.

Performance declined for all affective messages compared to three-way classification. Table 3 shows the confusion matrix that resulted when the adult-directed utterances were added: The percent of correctly classified approvals dropped from 80%

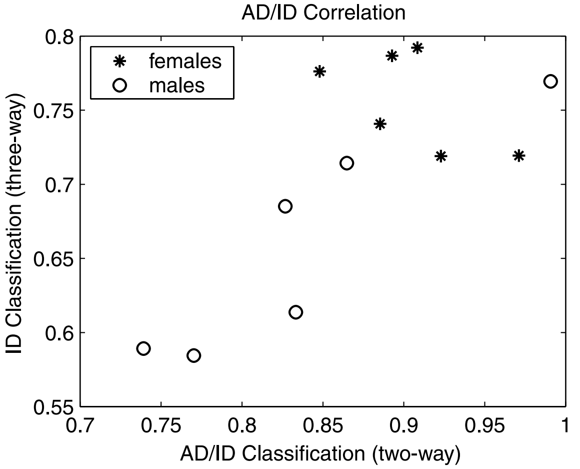

Fig. 13. Speaker-dependent classification results plotted as a function of adult-directed versus infant-directed (horizontal axis) affective message. In the interest of clarity, error bars have been hidden. Average vertical error bar is ±0.12; average horizontal error bar is ±0.09.

to 62%; that of prohibitions dropped from 42% to 23%. The neutral or adult-directed utterances were identified with the same accuracy as were the approval messages.

### 5.7. Adult-directed versus infant-directed correlation

Fig. 13 shows the correlation between our ability to discriminate infant-directed from adult-directed speech and our ability to recognize the affective message in infant-directed speech. Our hypothesis was that some speakers would be better than other speakers at making, or would be more willing to make, the acoustic gestures that characterize infant-directed speech. Fig. 13 shows a strong correlation, $r = 0.74$, between our classifiers' ability to recognize an utterance as infant-directed and their ability to recognize the affective message.

### 5.8. Speaker-dependent models

Psychologists and infant-development researchers have looked for a set of features that allow classifiers to operate optimally across speakers. However, emotional responses, and people's willingness to display them, vary widely across individuals, as may the way that people convey an affective message using speech, or in different situations. Thus, speaker-dependent classifiers should be more accurate than speaker-independent classifiers. Fig. 14 shows our speaker-dependent classification results. Except for the analysis of the utterances from one male speaker, our speaker-dependent classifiers were more accurate than were our speaker-independent ones.

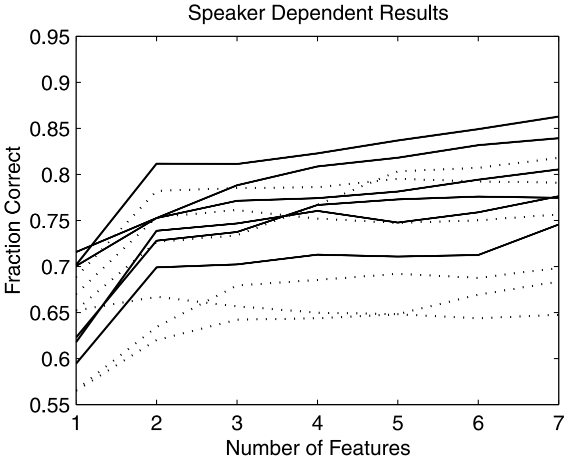

Fig. 14. Speaker-dependent results in a three-way (approval, attention, prohibition) classification test. The results from male speakers are shown as dotted lines and female speakers are shown with solid lines.

Utterances of individual speakers were classified correctly 65–87% of the time in this three-way test. The results are nearly all above the average score shown in Fig. 4, demonstrating that knowing the identity and the particular style of a speaker is important for classifying his or her utterances accurately, at least when classifying on the basis of our features.

## 6. Summary

BabyEars uses simple acoustic features to recognize affective messages. Using models that have only five to seven acoustic features, it classified correctly 65–75% of a large collection of infant-directed adult utterances.

Although it is difficult to make comparisons across different human listeners, computers, and tasks, we are encouraged by the accuracy of our recognizers compared to that of both our own and previously reported human listeners. For example, consider that our listeners had access to the words and the prosody, yet were not in complete agreement regarding how to classify the strongest utterances. Our adult listeners agreed with 96% of the classifications that we assigned initially during segmentation, which we assumed to be 100% accurate.

Our results are comparable to those reported by other investigators who controlled the linguistic message. Engberg and his colleagues (Engberg et al., 1997) asked an actor to read the same passage with five different emotions: neutral, surprise, happiness, sadness and anger. Their human listeners were able to judge the affective message of the same set of words 67% of the time in this five-way test. Roy (Roy and Pentland, 1996) played utterances backwards (to hide the linguistic message) to listeners who correctly classified 65–75% as approving or disapproving. As described in Section 2, Fernald's listeners classified 60–80% of low-pass filtered utterances correctly in a five-way test.

BabyEar's classifications were more accurate for female utterances than for male utterances. This result is surprising, insofar as adult listeners judged the utterances' strengths to be similar. As described in Section 5.5, there are many reasons why our classifiers might work differently on utterances made by male versus female speakers. Nonetheless, our results provide strong evidence that male and female speakers used different features to convey the same affective messages to their infants.

Global features allowed BabyEars to perform well, reducing the need to segment precisely the incoming audio. This result will become less important as speech-recognition systems improve to the point where they can make accurate judgments about utterance boundaries.

Our speaker-dependent recognizers performed more accurately than did our speaker-independent recognizers. We did not collect sufficient data to decide whether this performance difference was due to limitations of our classifiers, or whether affective vocal messages are more understandable if you know the prosodic customs of the speaker.

One issue not addressed by this study is the semantic versus prosodic content of our test material. The infants and the BabyEar's classifier were only listening to the prosodic signal. We do not know how the adult speakers were splitting their message between the two channels; nor do we know how our adult listeners were making their judgements. At a reviewer's suggestion, we performed a small pilot experiment asking adults to rate the strength of the affective message in text transcripts of our speech database. We compared the acoustic and semantic strength ratings, but did not see any clear patterns as a function of affective class, gender of the speaker, or prosodic strength.

We used a large collection of infant-directed utterances to judge BabyEar's performance in recognizing affective vocalizations. We found that a small handful of features is sufficient to allow us to perform this task at near-human accuracy. Building BabyEars is one step toward building machines that understand the emotional messages communicated by humans.

## Acknowledgements

Many people helped us perform this study. Jocelyn Scheirer, John Pinto, Helen Shwe and Jennifer Orton collected the data, with technical assistance from Kris Force and Dan Levitin. Jennifer Smith and Eric Scheirer helped us to do the statistical analysis. Discussions with Anne Fernald and Dave Huron gave us ideas for this work. Gaile Gordon and Michele Covell provided support and counsel. Lyn Dupré edited the manuscript. Detailed suggestions from a number of reviewers have greatly improved this manuscript.

## Appendix A

We made eight kinds of low-level acoustic measurements on each utterance. For each feature, we calculated the value of this statistic over each of the first, middle and last third (as shown in Fig. 1) of the utterance and over the entire utterance.

*Pitch variance: The variance of all valid pitch estimates.* The variance has units of octaves squared. If the utterance is all unvoiced, or no pitch can be determined, then the pitch variance is zero.

*Pitch slope: The linear regression of valid pitch data.* The slope has units of octaves per frame (20 ms). If the utterance is all unvoiced, or no pitch can be determined, then the pitch slope is zero.

*Pitch range: The difference, in octaves, between the highest and the lowest pitch values in the utterance.* If the utterance is all unvoiced, or no pitch can be determined, then the pitch range is zero.

*Mean pitch: The mean pitch of the utterance.* If the utterance is all unvoiced, or no pitch can be determined, then the mean pitch is zero. This feature, by itself, does not allow BabyEar's to judge affect, but we included it because knowledge of whether the speaker is male or female might make other judgments more accurate.

*Mean delta pitch: The change in pitch between frames (20 ms).* We calculated the difference in pitch for every pair of adjacent frames for which we had valid pitch estimates. To avoid problems with octave errors, we ignored frames in which the pitch changed by more than 0.75 octaves. The units of this measure are in octaves per frame. This measure is similar to the mean slope measure.

*Absolute value of the mean delta pitch: The total of the absolute value of the change in pitch from frame to frame.* The pitch slope and mean-delta pitch measures do not give any information if the pitch rises and then falls by the same mount. This measure finds the absolute value of the delta pitch before computing the mean. Thus, larger pitch excursions, no matter what their direction, should produce a larger result than do gradual pitch glides. Again, to avoid problems with octave errors, we ignored frames in which the pitch changed by more than 0.75 octaves.

*Delta MFCC: The total change in the MFCC coefficients over the segment.* A 13th order MFCC calculation was made at each frame. The energy term, C0, was ignored. The sum of the squares of the frame-to-frame differences between the 12 remaining coefficients was calculated. We calculate the mean over the entire utterance. Well-articulated utterances have larger excursions in spectral or formant space and give higher values for this measure. The variance of this measure, which indicates how fast the spectral transitions were, is also interesting, but was not used in this study.

*Energy variance: The variance of the frame-by-frame energy.* This measure is amplitude independent; an overall gain change will offset the entire utterance (in dB), but will not change the variance.

## Appendix B

Many features allowed us to build good classifiers. As shown in Fig. 2, different sets of features produced nearly identical results. In this appendix, we summarize the features used in each study in this paper, from first chosen to last chosen. We use the following abbreviations for each feature name: pitch variance (PV), pitch slope (PS), pitch range (PR), mean pitch (MP), mean delta pitch (MDP), mean absolute value delta pitch (MADP), delta MFCC (MFCC), and energy variance (EV). We use suffixes to indicate the time period (1 = first, 2 = second, or 3 = third third, and g = global).

Fig. 4: Strength plot. All: PSg, PRg, MPg, MADPg, MFCC3, PV2, MFCC2. Strong: MFCCg, PSg, EV1, PS3, MP2, PVg, MFCC3. Very Strong: MFCCg, PSg, MP3, EVg, MP2, PRg, MDP3.

Fig. 5: Global/local plot. All: PSg, PRg, MPg, MADPg, MFCC3, PV2, MFCC2. Global: PSg, MPg, PRg, MADPg, MFCCg, EVg, MDPg.

Fig. 6: Message plot. All tests: PSg, PRg, MPg, MADPg, MFCC3, PV2, MFCC2.

Fig. 9: Gender plot. All: PSg, PRg, MPg, MADPg, MFCC3, PV2, MFCC2. Male: PSg, MPg, PRg, MFCC3, EV2, MP1, MADPg. Female: PSg, PVg, PS1, MP3, MDP2, PR3, EV1.

Fig. 11: AD/ID discrimination. All: PS2, MFCC3, MFCCg, MDP3, MPg, EVg, MDP1. Female: PS2, PS3, PR3, MPg, MFCC3, EV3, MDP3. Male: MPg, PSg, MFCC3, MP3, PS3, MFCC2, EV2.

Fig. 12: Four-way classification. All: PSg, MP1, PVg, MFCC3, PS3, MFCC2, EV1. Female: PS2, PS3, PR3, PS1, MPg, MFCC3, PVg. Male: MPg, MFCC1, MFCC3, MDPg, MADP3, PS2, PS1.

Fig. 14: Speaker dependent results. (1) MFCC2, PRg, MADP2, MADP3, MDP2, PV1, MP3. (2)

EV1, PRg, PR1, MDP1, PR3, EVg, MPg. (3) MPg, MP2, MFCC3, PR2, MDPg, PVg, PV2. (4) MP1, MFCC2, MP2, PR3, PVg, EVg, PS1. (5) MDPg, PRg, EV3, MFCCg, PVg, MP1, MPg. (6) MFCCg, PR1, EVg, MFCC1, MADP3, MDPg, EV1. (7) PSg, MP2, MDP3, PVg, PS3, PV1, MP1. (8) PS3, MADPg, MP2, MADP3, MDP3, MDPg, MFCC2. (9) MP1, PSg, MFCCg, MPg, MDPg, PRg, PV2. (10) MP1, PVg, MDP3, EVg, PSg, MADP1, MP3. (11) MP1, PV3, MFCCg, PR1, MDPg, PS1, PS2. (12) PS3, MFCC2, EV1, EV3, MFCC3, MADP3, PRg.

## References

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK.

Droppo, J., Acero, A., 1998. Maximum a Posteriori Pitch Tracking. In: Proc. Internat. Conf. on Spoken Language Processing. Sydney, Australia. December 1998.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P. 1997. Design, recording and verification of a Danish emotional speech database. In: Proc. EuroSpeech '97, Rhodes Greece, Vol. 4, pp. 1695–1698.

Fairbanks, G., Pronovost, W., 1939. An experimental study of the pitch characteristics of the voice during expression of emotion. Speech Monographs 6, 87–104.

Ferguson, C.A., 1964. Babytalk in six languages. American Anthropologist 66 (part 2), 103–114.

Fernald, A., 1985. Four-month-old infants prefer to listen to motherese. Infant Behavior and Development 8, 181–195.

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., 1992. Human maternal vocalizations to infants as biologically relevant signals: an evolutionary perspective. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford.

Fernald, A., 1993. Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. Developmental Psychology 64, 657–674.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., De Boysson-Bardies, B., 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. Journal of Child Language 16, 477–501.

Garnica, O., 1977. Some prosodic and paralinguistic features of speech to young children. In: Snow, S.E., Ferguson, C.A. (Eds.), Talking to Children: Language Input and Acquisition. Cambridge University Press, Cambridge.

Hunt, M.J., Lennig, M., Mermelstein, P., 1980. Experiments in syllable-based recognition of continuous speech. In: Proc. 1980 ICASSP, Denver, CO, 1980, pp. 880–883.

Katz, G.S., Cohn, J.F., Moore, C.A., 1996. A combination of vocal f0, dynamic, and summary features discriminates between three pragmatic categories of infant-directed speech. Child Development 67, 205–217.

Lamel, L.F., Rabiner, L.R., Rosenberg, A.E., Wilpon, J.G., 1981. An improved endpoint detector for isolated word recognition. IEEE Trans. ASSP, Vol. ASSP-29. pp. 777–785.

McRoberts, G., Fernald, A., Moses, L., under review. An acoustic study of prosodic form-function relations in infant-directed speech: Cross-language similarities. Developmental Psychology.

Nabney, I., 2001. NETLAB: Algorithms for pattern recognition. Springer, London, UK.

Papousek, M., Papousek, H., 1991. The meanings of melodies in motherese in tone and stress languages. Infant Behavior and Development 14, 415–440.

Picard, R.W., 1997. Affective Computing. MIT Press, Cambridge MA.

Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., Fong, C., 1991. The use of prosody in syntactic disambiguation. Journal of the Acoustical Society of America 90, 2956–2970.

Roy, D., Pentland, A., 1996. Automatic spoken affect classification and analysis. In: IEEE Face and Gesture Conference, Killington. VT. pp. 363–367.

Sherer, K.R., 1986. Vocal affect expression: A review and a model for future research. Psychology Bulletin 99 (2), 143–165.

Skinner, E.R., 1935. A calibrated recording and analysis of the pitch, force, and quality of vocal tones expressing happiness and sadness. Speech Monographs 2, 81–137.

Slaney, M., McRoberts, G., 1998. Baby Ears: a recognition system for affective vocalizations. In: Proc. 1998 Internat. Conf. Acoust., Speech Signal Process., Seattle, WA. Also available at http://rvl4.ecn.purdue.edu/~malcolm/interval/1997-063.

Stern, D., Speiker, S., McKain, K., 1982. Intonation contours as signals in maternal speech to prelinguistic infants. Developmental Psychology 18, 727–735.

Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). In: Klein, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier Science, Amsterdam, pp. 495–518.

Williams, C.E., Stevens, K.N., 1972. Emotions and speech: some acoustic correlates. Journal of the Acoustical Society of America 52, 1238–1250.