This is a reprint of a paper that was published in the *Proceeding of the 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, WA, May 12-15, 1998.

BABY EARS: A RECOGNITION SYSTEM FOR AFFECTIVE VOCALIZATIONS

Malcolm Slaney and Gerald McRoberts[1]

http://www.interval.com/papers/1997-063/

Interval Research Corporation
1801 Page Mill Road, Building C
Palo Alto, CA 94304, USA

ABSTRACT

We collected more than 500 utterances from adults talking to their infants. We automatically classified 65% of the strongest utterances correctly as approval, attentional bids, or prohibition. We used several pitch and formant measures, and a multidimensional Gaussian mixture-model discriminator to perform this task. As previous studies have shown, changes in pitch are an important cue for affective messages; we found that timbre or cepstral coefficients are also important. The utterances of female speakers, in this test, were easier to classify than were those of male speakers. We hope this research will allow us to build machines that sense the "emotional state" of a user.

1 VOCAL AFFECT

The goal of a new field of study known as *affective computing* is to design machines that understand and respond to human emotions [9]. There is a range of information available for the development of a human–machine interface based on emotion. Locally, we can monitor physiological measures of human emotional state, or we can judge at a distance using visual or vocal expressions of emotion. In the latter approach, we attempt to relate aspects of speech *prosody* (e.g., variations in the pitch, rhythm, and loudness of speech) to the affective state, or pragmatic intent, of the speaker.

Studying spontaneous emotion is difficult. Much of the engineering work on vocal expressions of emotion is based on actors reading sentences in specified emotional tones [2, 8, 11]. Such expressions at best merely resemble real emotional expressions. We need a way to capture affective and pragmatic vocalizations that are both spontaneous and clearly identifiable.

A promising solution to this problem is to use parents' speech to infants. 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.

Not only is the prosodic message clear, but it may also be universal. Fernald and her colleagues have shown that the prosodic patterns parents use to convey affective and pragmatic messages such as prohibition, praise and attention-bid are similar across languages and that infants respond appropriately to these vocalizations even in unfamiliar languages [3, 4, 7].

Normally, the words and prosody of an utterance contribute to both the linguistic and affective message. We want to see how much of the affective message can be recovered from simple acoustic measures of the speech signal. Note that infants and current speech-recognition systems operate on different aspects of the vocal signal. Infants understand the prosodic message conveyed by "gooood giirrrllll" and "NO! STOP!" long before they understand the words. Speech-recognition systems worry about the words and mostly ignore the prosody.

Thus, we simplify: We study how adults convey affective messages to infants using prosody. We do not attempt to recognize the words, let alone to distill more nebulous concepts such as satire or irony. We analyze speech with low-level acoustic features and discriminate approval, attentional bids, and prohibitions from adults speaking to their infants. We built automatic classifiers to create a system, Baby Ears, that performs the task that comes so naturally to infants. We believe that adult-directed speech contains the same affective messages as the speech we studied, with the same prosodic patterns, although attenuated.

The remainder of this paper describes our data collection (Section 2), signal-processing techniques (Section 3), and results (Section 4).

2 DATA COLLECTION

We collected two kinds of experimental data. In the primary experiment, we collected acoustic data from parents talking to their infants. In the second experiment, different adult listeners judged whether each utterance was best classified as an approval, attentional bid, or prohibition, and judged the strength of the message.

2.1 Acoustic Data

We recorded 12 parents—six mothers and six fathers—talking to their 10- to-18-month-old infants in a quiet room. Each recording session lasted about 1 hour, during which the parents were asked to play and interact normally with their child. Several toys were placed in the room. We asked the parents to use 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.

We recorded audio from the parent, using a wireless microphone mounted on a lightweight headset, directly onto a computer's hard disk and then downsampled to 22kHz.

A trained experimenter segmented the recorded audio into discrete sentences and classified the utterances into three classes: approval, attention, and prohibition. Typical examples of each category follow; note that these words do not do justice to the prosodic contours:

- 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 30 to 50 utterances, each comprising one phrase or sentence. We analyzed 212 approval utterances, 149 attentional bids, and 148 prohibitions.

1. Current Address: Gerald McRoberts, Department of Psychology, 17 Memorial Drive E., Lehigh University, Bethlehem, PA 18015.

2.2 Subjective Classifications

In a separate test, we had seven adult subjects listen to each of the segmented utterances and judge the utterance's category and strength. The adult listeners had no training in either linguistics or psychology. None of the listeners were familiar with our hypothesis or method. Each listener rated each utterance as an approval, attentional bid, or prohibition, and assigned it a strength on an arbitrary scale from 1 to 5.

The utterances were grouped into three sets according to the results of the listener test:

- All data: All utterances, including those for which the listeners did not agree with the original classifications
- Strong data: Utterances for which 5 out of 7 listeners agreed with the initial classification and the average strength was above 2.5
- Very strong data: A subset of strong data, with an average strength above 3.0

We report results on these three sets of data, in addition to gender- and subject-dependent tests.

3 ANALYSIS

We analyzed the speech using three classes of features: pitch, formant transitions, and energy variations. In brief, we postulated that speech that had long, smoothly varying sounds would indicate approval, whereas sounds that changed quickly would be attention bids or prohibitions. Several variations of these parameters were measured and analyzed for their ability to classify these utterances.

We performed signal processing on each utterance, and built multidimensional classifiers to perform the classification experiments.

3.1 Signal Processing

Each utterance was processed automatically with a frame rate of 50Hz. A speech–silence discriminator segmented each utterance at phrase boundaries [6]. We then chose the longest phrase in each utterance for additional processing.

For analysis, we processed each utterance as a whole, and split each utterance into three segments: the first, middle, and final third of the sound. Thus, for each feature—for example, the pitch range—we had four measurements over different time periods.

Three kinds of analysis were done on each temporal period of each utterance: pitch, cepstral or formant changes, and energy.

We analyzed the pitch of each utterance using a high-quality dynamic-programming algorithm [12]. The pitch module produced estimates of the speech signal's pitch, measured in Hertz. We then computed the log, base 2, of this number to collapse the pitch estimate into octaves and to put the measurement on a perceptual scale. We did 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.

We measured several statistics related to the pitch: the variance, slope, range (maximum minus minimum), and mean. We also measured 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.

We used mel-frequency cepstral coefficients (MFCC) [5] to measure the formant information in the speech. MFCC parameters are often used in speech recognition as a simple measure of what is being said. We wanted to investigate whether the speed with which these parameters changed would be a useful feature. Thus, we measured the mean frame-by-frame change in the MFCC parameters during each segment of the utterance. In this calculation, we ignored the energy, or C0 component, and summed the absolute value of the changes in the remaining coefficients.

Finally, we also computed the variance of the energy in dB in each frame, across each utterance.

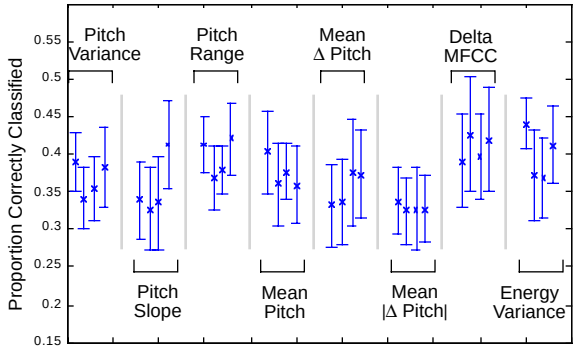

Figure 1. Three-way classification results for all our data with one feature. Bracketed points show the fraction of correct classifications with the indicated measure. The four points represent, from left to right, the first, second, and final third of each utterance, and global measurements. Error bars show ±1 standard deviations of the individual bootstrap error estimates.

3.2 Classification

We built many multidimensional discriminators to put each utterance into the proper class. We judged Baby Ears' performance based on whether the automatic classifier produced the same label as the experimenter.

We used a Gaussian mixture model, GMM [10], with 10 Gaussians per class, to model each class of data. Since we had a limited set of data, we used the .632 bootstrapping procedure [1] to estimate our performance. We trained with a set of data, chosen randomly with replacement, equal in size to the original data set. We tested our classifiers with all data that were not used in training. We repeated this task 100 times per discriminator, then averaged the results to find an estimate of the mean and standard deviation of the recognizer's performance with that set of features. We obtained similar results with optimal linear discriminators.

Finally, we built an optimal classifier using greedy selection. At each step, we trained three GMMs, one for each class, with the current set 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 *n* best features for making this classification. This test gave us information about which features were adding the most information to the decision.

4 RESULTS

The seven adult listeners agreed unanimously with our initial classifications in 79% of the examples. There were 430 utterances classified consistently by 5 out of 7 listeners with an average strength greater than 2.5, and 318 utterances with an average strength greater than 3.0.

Weak affective utterances in our database, as defined by our listener's strength measurements, often combined a strong linguistic message with a different prosodic message. For example, "Nicholas, don't do that" said with a soft, pleading voice is a linguistic prohibition said with an encouraging (or perhaps resigned) affective message.

Figure 1 shows the classification results for individual acoustic features. Not one of the features, by itself, allows the classification to be made with accuracy much greater than chance (33%). The

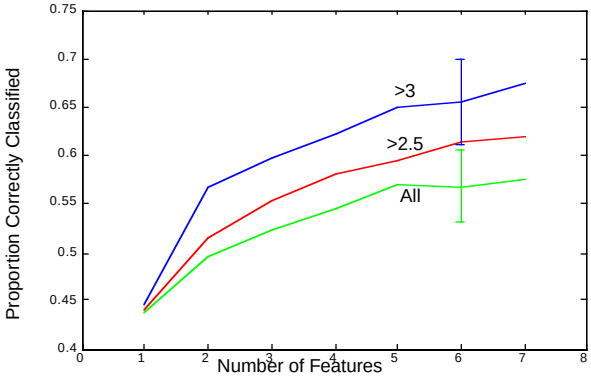

Figure 2. Performance as features are added for three different sets of data, as described in Section 2.2. Error bars show ±1 standard deviation for two representative bootstraps measurements. Other estimates have similar variance.

pitch range and energy variance look most promising in this simple test.

Figure 2 shows classification results for all speakers as we add more features to the classifier. Classification performance increases as more features are added, then levels off above 57% with five to seven features.

Classification performance improves when we consider only those utterances that are strong and unambiguous. Figure 2 compares classification performance when training included all utterances, only those utterances that had an average strength rating greater than 2.5, or only those utterances with average ratings greater than 3.0. Classification results are higher when the data set is limited to vocalizations with the highest strength ratings.

For reasons that we do not understand, classification performance was higher for female speakers than for male speakers (see Figure 3.) The female utterances were classified at a rate up to 67% correct, whereas male utterances were classified correctly 57% of the time. This difference could be caused by four factors. First, the acoustic features that we analyzed may not be optimal for male speech. Second, our procedure may not have captured the full affective range of the male speakers. Third, female speakers may be more skillful or more practiced in producing characteristic infant-directed speech. Fourth, male speakers may have been less willing or able in our corporate laboratory environment to produce the prototypical utterances of infant-directed speech. Average strength ratings by our listeners were equal for male and female speakers.

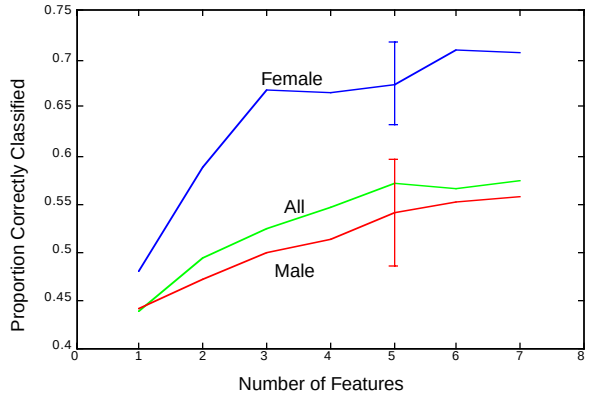

Figure 3. Classifier performance for all utterances (middle line), versus males only (bottom line) and females only (top line.)

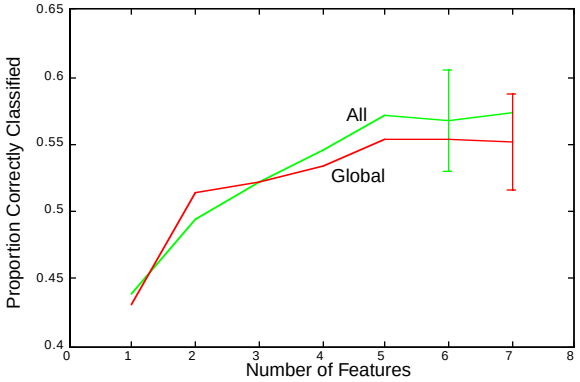

Figure 4. Three way classification results with all features (top curve) versus the global features only.

The results in Figures 1–3 are based on the segmentation approach used by McRoberts [7]. Unfortunately, segmenting speech into discrete utterances, so that they can be split accurately into thirds, is difficult. We avoid this problem if we use only global features to make a classification. These features are less sensitive to how well the speech is segmented. Figure 4 shows our results using only global features. Our recognition rates were slightly lower than when we use the full feature set.

Psychologists and infant researchers have looked for a set of features that operate optimally across speakers. However, emotional responses, and people's willingness to display them, vary widely among individuals. The way that people convey an affective message often varies across speakers, and even across situations. Thus speaker-dependent classifiers should work better than speaker-independent classifiers. Figure 5 shows our speaker-dependent classification results. Except for two speakers, both male, our results are much better with speaker-dependent classifiers. Training and testing speaker-independent classifiers with all utterances from the nine best speakers raises our classification performance to 66%.

Figure 6 shows plots of the decision surface for one of our best classifiers (recognizing female utterances with a strength greater than 3.0). These decision surfaces are plotted as a function of two different sets of variables.

Many features allow us to build good classifiers. The most commonly chosen feature is global pitch range, followed by global MFCC, global pitch slope and variance of the energy, in the first

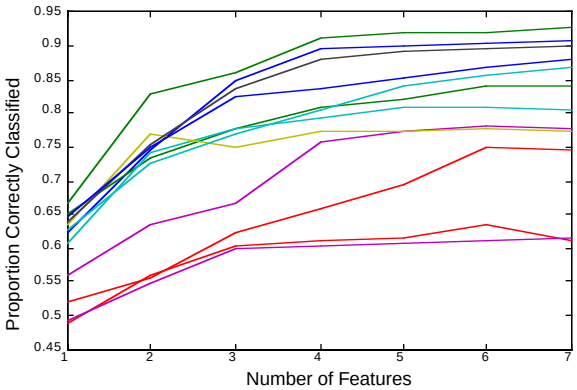

Figure 5. Performance on classification task as features are added for each of 12 adults speakers. The three speakers with the worst classification performance are male.

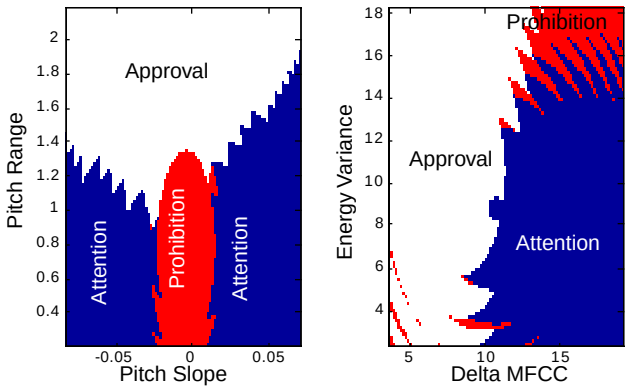

Figure 6. Bivariate discrimination surfaces for two different sets of global variables.

segment. Using these features, our classifiers demonstrated a recognition rate of 53, 54, and 58% correct on the three sets of data in Figure 2. Using just global MFCC and global pitch range, they obtained 51, 54, and 56% correct.

We are encouraged by the accuracy of our recognizers compared to that of human listeners. While performance was not as high as our adult listeners, who had access to the linguistic message, it is comparable to results reported by others who controlled the linguistic message. For example, Engberg [2] reported listeners' accuracy at 65% when judging emotional messages in Danish in which the linguistic message was controlled by using the same sentence for different affective messages. Other studies have tried to mask the linguistic message through filtering [4, 11], but this often introduces artifacts which may alter the affective message.

5 CONCLUSIONS

Baby Ears is a system that uses simple acoustic features to recognize affective messages with the same accuracy as that of some human tests.

We found it easier to classify female then male utterances from our database. This result is surprising, because adult listeners judged the utterances' strengths to be similar. We do not know whether male and female affective vocalizations are different in some interesting way, or whether our data are not representative of real-life situations.

Global features perform well in our classifiers, reducing the need to segment precisely the incoming audio. This result will become less important as speech-recognition systems improve, eventually allowing good sentence boundaries to be judged. Our work did not consider any melodic contour matching. That approach would work best with good segment boundaries, and perhaps with linguistic information to guide the pattern recognition.

Speaker-dependent recognizers perform more accurately than do speaker-independent recognizers. We do not have sufficient data to decide whether this performance difference is due to limitations of our classifiers, or whether affective vocal messages are more understandable if you know the prosodic customs of the speaker.

We used a large collection of infant-directed utterances to judge our results. We found that a small handful of features is sufficient to perform this task, at near-human levels. Baby Ears is a large step towards building machines that understand the emotional messages communicated by humans.

6 ACKNOWLEDGEMENTS

Many people helped to make this study possible. Jocelyn Riseberg, 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 with the statistical analysis. Discussions with Anne Fernald and Dave Huron helped to motivate this work. Gaile Gordon and Michele Covell provided support and counsel. Lyn Dupré provided editorial assistance.

7 REFERENCES

[1] Brad Efron, Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[2] Inger S. Engberg, Anya V. Hansen, Ove Andersen, Paul Dalsgaard. "Design, recording and verification of a Danish emotional speech database." *Proceedings of EuroSpeech '97*, Rhodes Greece, Vol. 4, pp.1695–1698, 1997.

[3] A. Fernald. "Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages." *Developmental Psychology*, Vol. 64, pp 657–674, 1993.

[4] A. Fernald. "Intonation and communicative intent in mother's speech to infants: Is the melody the message?" *Child Developmen*t, Vol. 60, pp. 1497–1510, 1989.

[5] M. J. Hunt, M. Lennig, P. Mermelstein. "Experiments in syllable-based recognition of continuous speech." *Proceedings of 1980 ICASSP*, Denver, CO, pp. 880-883, 1980.

[6] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon. "An improved endpoint detector for isolated word recognition." *IEEE Transactions on ASSP.* Vol. ASSP-29, pp. 777–785, August 1981.

[7] Gerald McRoberts, Anne Fernald, Lou Moses, "An acoustic study of prosodic form-function relations in infant-directed speech: Cross language similarities," *Development Psychology*, (in press).

[8] J. E. H. Noad, S. P. Whiteside, P. D. Green. "A macroscopic analysis of an emotional speech corpus." *Proceedings of EuroSpeech '97*, Rhodes, Greece, Vol. 1, pp. 517–520, 1997.

[9] Rosalind W. Picard. *Affective Computing.* MIT Press, Cambridge, MA, 1997.

[10] R. A. Redner, H. F. Walker, "Mixture densities, maximum likelihood, and the EM algorithm," *SIAM Review*, Vol. 26, pp 195-239, 1984.

[11] Deb Roy, Alex Pentland, "Automatic spoken affect classification and analysis," *IEEE Face and Gesture Conference*, Killington, VT, pp. 363–367, 1996.

[12] David Talkin. "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleign and K. K. Paliwal, eds., Elsevier Science, Amsterdam, pp. 495–518, 1995.

APPENDIX

We summarize the features used in each figure, 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 subscripts to indicate the time period.

Figure 2:
All: EV_1, PR_g, MDP_g, $MFCC_g$, $MFCC_1$, EV_2, PS_g
> 2.5: EV_1, PR_g, $MFCC_g$, MDP_3, $MFCC_2$, MP_3, $MFCC_1$
> 3: $MFCC_g$, PR_g, MP3, EV_1, MP_1, MDP_g, $MFCC_1$

Figure 3:
All: EV_1, PR_g, MDP_g, $MFCC_g$, $MFCC_1$, EV_2, PS_g
Male: PS_g, MP_1, PR_1, $MFCC_3$, MDP_g, PR_g, MDP_1
Female: PR_g, PS_g, MP_g, PV_g, MFCC2, EV_1, PS_1

Figure 4:
All: EV_1, PR_g, MDP_g, $MFCC_g$, $MFCC_1$, EV_2, PS_g
Global: PR_g, $MFCC_g$, EV_g, PS_g, MP_g, $MADP_g$, PV_g