Proceedings of 2004 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 28 - October 2, 2004, Sendai, Japan

Prosody Recognition in Male Infant-Directed Speech

Avram Lev Robinson-Mosher and Brian Scassellati
Department of Computer Science
Yale University
New Haven, CT, USA
avram.robinson-mosher@aya.yale.edu and scaz@cs.yale.edu

Abstract— **Robots designed to learn from and interact with humans require an intuitive method for humans to communicate with them. Normal human speech is very difficult to process, requiring many kinds of complex analysis for robots to interpret it. An intermediate method for communication is recognition of prosody, the affective content of speech. Using prosody recognition, a human interacting with a robot can reward or punish its actions by scolding or praising it. In this project, prosody recognition of male voices is performed by feature-based analysis of sound files containing short utterances, which were recorded from subjects who were directed to emulate infant-directed speech, which generally contains exaggerated prosody [1]. The features used are extracted from the energy and pitch contours in the preprocessing stage. The classifier discriminates amongst four affective classes of speech and neutral utterances. The four classes are prohibition, attentional bids, approval, and soothing, while the neutral utterances are speech which carries none of the above affective intents. Discrimination is performed using a multi-stage k-nearest neighbor classifier. The five-way single-stage classifier operates at 62.5 accuracy on the entire male speech data set, while the female single-stage classifier classifies 66.7 percent correctly. Chi-square analysis resulted in a p of less than or equal to 0.001 for each. The data seem to indicate that while female voice data may be somewhat easier to classify than male, fundamental differences that make male utterances unsuitable for classification do not exist.**

Keywords; prosody; speech-recognition; human-computer interaction

I. Introduction

Studies have shown [4] that infants respond to the prosodic content of speech before they are capable of processing its linguistic content. Giving a robot the capacity for prosody recognition allows human caregivers to communicate intuitively with it. Functional prosody recognition for female robot-directed speech has been achieved by the Kismet group at MIT [1]. The motivation of our project was the desire to extend the work done on Kismet and to create a prosody recognizer that would function on all robot-directed speech, regardless of the speaker, and implement that recognizer on Nico, a robot meant to emulate the behavior of a nine month old child, currently in development at the Yale Social Robotics laboratory. This paper details the development of a male prosody recognizer and its comparison with a female prosody recognizer on data collected in our experimental environment.

Since humans can generally recognize the affective intent of other humans without linguistic information, it should be possible to train a computer to recognize affective intent as well. Fernald has done work identifying pitch contours that are characteristic of approval, prohibition, attention, and comforting speech [4], and the Kismet prosody work gives details on methods for attempting such classification. Given this, creating a working speaker-independent prosody classifier designed to operate on the kind of exaggerated prosody that adults generally direct towards infants should be feasible, and is desirable for the purpose of allowing humans to train a robot interactively. This is especially true in the context of social robotics, where the recognizers do not act in a vacuum. In an interactive system, while the recognizer is trained on speech data, speakers are also trained by the robot based on how it reacts to their speech.

Slaney has also done work on infant-directed prosody produced by male and female adult speakers [6]. The chief difference in this study is that the prosodic utterances were explicitly partitioned into male and female data sets. Other studies have incorporated prosodic information as well, some for its use in general speech analysis, some for direct analysis [7].

Prosodic information can play an important part in most areas of speech processing and synthesis. Studying prosody in infant-directed speech, where it is most exaggerated, can provide an opportunity to learn fundamental truths about it that can aid in dealing with it in its subtler forms in dialogue amongst adults. Analysis of these early stages of interaction may also lead to interesting revelations about child development and the differences or similarities between male and female interactions with children.

II. Methods

A. Data Gathering and Processing

Subjects were placed in front of a workstation and instructed to provide samples of the five prosodic types (see Table I). In early data gathering subjects were supplied with situation or action cues to which they were asked to respond, along the lines of "Your child is crying"

0-7803-8463-6/04/$20.00 ©2004 IEEE

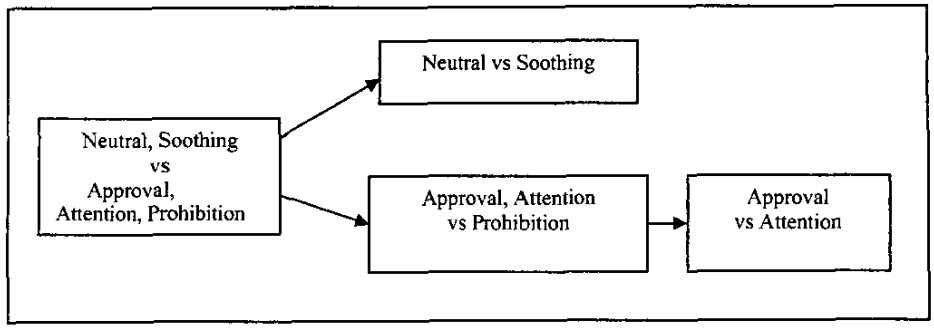

Figure 1. The subclassifier hierarchy of the multistage classifier

or "The dog just peed on the carpet." Subjects displayed difficulty coming up with statements or showing prosody in these scenarios. In later trials subjects were provided with a transcript of utterances from previous recordings from which they could choose if they desired, and the situation cues were removed. Each subject was asked to provide approximately twenty utterances from each of the four classes and neutral speech.

All data collection was performed in the same environment that the Nico robot is intended to operate in. The lab environment is a 20 by 20 foot room with high ceilings. Noise is limited to air conditioning and a server rack across the room from the recording setup.

Data was collected using the Matlab wavrecord function, a lapel microphone with preamplifier, and a Soundblaster Live! sound card on a PC running Windows 2000. The audio signals were recorded at 48800 Hz and stored as dual-channel WAVE files. The raw data was converted into pitch, energy and first-formant contours using Edinburgh Speech Tools [2] and Speech Filing System [3] library functions, and then preprocessing and feature extraction were performed in Matlab using those contours.

The preprocessing was chiefly performed to remove some noise and repair pitch halving and doubling errors by making assumptions about the normal pitch ranges and fitting curves to adjacent sets of points. The Edinburgh Speech Tools library performs voicedness determination, which along with the energy contour was used to break utterances up into segments, often corresponding to syllables. Several of the features used in the classifier operate on the individual segments, especially those based on matching certain shapes in the pitch contour such as the bell-shaped contour feature (F13) from [1].

B. *Feature Selection and Classification*

The k-nearest neighbor classification method was chosen due to comparable or superior performance to other learning methods tested and faster trial speed, which facilitated testing. Other classification methods tested include Gaussian Mixture Models, simple neural networks with varying number of hidden nodes, self-organizing maps, and principle components analysis. In an attempt to gain some scalability, the number of neighbors used in the classifier was scaled by the total number of samples. Data were normalized by mean and variance before neighbor distance calculations were performed. Following Cynthia Breazeal's work [1], both single-stage and multistage classifiers were attempted. The class breakdown used was nearly identical to that of the Kismet multistage classifier (Fig. 1).

The Weka machine-learning package [5] was used to select the features for the k-nearest neighbor classifiers. All sets of features were selected by using Weka's genetic search function on cross-validated training sets over 10

TABLE II. SAMPLE UTTERANCES

Prosodic class	Utterance
Neutral	"Crabs taste bad." "There's a shelf on the wall."
Approval	"Wow, good boy!" "Good job!"
Prohibition	"No, don't do that." "Stop it!"
Attention	"Hey, over here!" "Look at this!"
Soothing	"Aw, feel better." "It'll be okay..."

TABLE I. FEATURES FOR CLASSIFIER

F1	Mean pitch	F17	Min segment length
F2	Pitch variance	F18	Max segment length
F3	Max pitch	F19	Avg. segment linear
F4	Min pitch	F20	Pitch energy mean
F5	Pitch range	F21	Pitch energy variance
F6	Delta pitch mean	F22	Pitch peak ratio
F7	Abs delta pitch mean	F23	Pitch error ratio
F8	Mean energy	F24	Pitch to energy ratio mean
F9	Energy variance	F25	Pitch to energy ratio variance
F10	Energy range	F26	Pitch delta by mean
F11	Max energy	F27	Max pitch slope
F12	Min energy	F28	Energy max range
F13	Max rise/fall	F29	Sigmoid fit segment
F14	Contour slope	F30	First formant mean
F15	Pitch segments	F31	First formant variance
F16	Avg. segment length	F32	First formant pitch ratio

TABLE III. CLASSIFIER PERFORMANCE

Classifier	Features	Percentage correct
Male (single-stage)	F1 F4 F5 F6 F8 F24 F25	62.5
Male (high-energy vs low-energy)	F8 F24	81.1
Male (neutral vs soothing)	F1 F4 F5 F8 F18 F22 F24 F26	84.3
Male (prohibition vs approval-attention)	F1 F2 F6 F10	86.3
Male (approval vs attention)	F1 F5 F7 F17 F19 F20 F24 F28	72.9
Male (multi-stage)		64.4
Female (single-stage)	F1 F3 F6 F7 F8 F9 F19 F22 F24	66.7
Female (high-energy vs low-energy)	F1 F24 F26	87.2
Female (netural vs soothing)	F1 F3 F10 F11 F16 F19 F22 F24 F32	78.6
Female (prohibition vs approval-attention)	F1 F6 F7 F8 F17 F19 F20 F24	83.2
Female (approval vs attention)	F1 F6 F7 F9 F12 F17 F22 F24 F26	91.1
Female (multi-stage)		75.1

folds. Features were chosen for inclusion in the final classifier based on how often they appeared in the final evolved classifier in Weka combined with their performance when added sequentially to the classifier. In order of Weka selection, each feature was tested for whether its addition improved the classifier and then included or not based on its performance.

This project used most of the features present in Breazeal's earlier work [1] along with some new features (Table II). Features F1 through F18 are from Breazeal's work. Feature 19 is the average across segments of the slope of the best-fit line. Features F20 and F21, pitch-energy mean and pitch-energy variance, refer to the pointwise product of the pitch and energy contours. Feature F22, pitch peak ratio, is the ratio of the peak in the last pitch segment to the highest peak in any previous pitch segment, and was included in an attempt to differentiate approval from attention. Feature F23, pitch error ratio, relates to how well pitch segments are fit by second or third degree polynomials and is only mentioned for completeness. Features F24 and F25, pitch-energy ratio mean and pitch-energy ratio variance, refer to the pointwise ratio of the pitch and energy contours. Feature F26, pitch delta by pitch mean, is a normalization of the absolute value of the delta pitch mean (F7) by the pitch mean (F1). Feature F27, max pitch slope, is the largest best-fit slope of any pitch segment, normalized by pitch mean. Feature F28, max energy range, is the range of energy maximums amongst pitch segments. Feature F29, the sigmoid fit segment, is an attempt to fit a pitch shape contour that seems characteristic of prohibition in male voices. Features F30 and F31, the first formant mean and first formant variance, are simple statistics on the first formant contour, while F32, first formant pitch ratio, is the mean of the ratio of the first formant to the fundamental frequency.

Features F20, F21, F24, and F25 in particular were added to explore any relevant time correlation that might exist between pitch and energy while still treating them in a

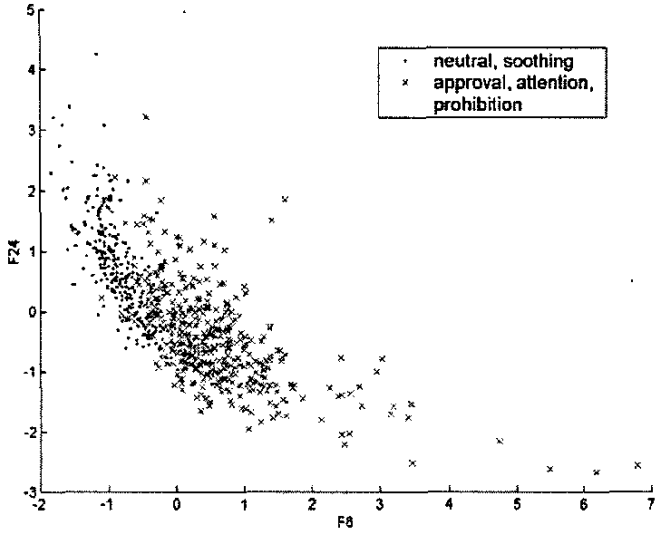

Figure 2. This feature pair was used to discriminate the high-energy vs low-energy subclassifier space in the male data set

general statistical manner rather than explicitly looking for predefined contours or relationships. Unlike some of the other features added, which were targeted at particular observed features of the pitch and energy contours, these features were added with only the thought that they explored possibilities missing in the current feature set. Along with F26, even if the time correlation were irrelevant they provided a set of operations on the existing features that could not be easily duplicated by a learning algorithm (i.e., if the ratio of F1 to F8 were significant on its own). The formant features were added when it seemed as if pitch and energy contours on their own might not provide sufficient complexity to fully capture prosodic variations.

III. RESULTS

The single-stage classifier achieved a mean performance of 62.5 percent with a standard deviation of 3.8 percent when tested on a set of 700 samples from six different male speakers, and the multistage classifier achieved 64.4 percent accuracy with a standard deviation of 3.6 percent. The single-stage classifier for female speakers achieved an average performance of 66.7 percent with a standard deviation of 5.6 percent over 1000 trials when tested on a set of 302 samples from three different female speakers, and the multistage classifier 75.1 percent accuracy with a 5.1 percent standard deviation. Chi-square analysis returned a p value of less than or equal to 0.001 for each of the single-stage classifiers. When run on a set of samples corresponding to only three male speakers (300 samples), the male single-stage classifier functioned at up to approximately 70 percent accuracy, depending on which speakers were used, while accuracy rates for the male multistage classifier tested on the same set of speakers fell within one standard deviation.

Fig. 1 contains details of the features used in each of the subclassifiers and the individual classifier performance results. All results are over 1000 trials except for the multistage classifier results, which are over 100 trials. Table III shows the performance of the single-stage male classifier, single-stage female classifier, and single-stage male classifier on a subset of 300 elements. Figures 2, 3 and 4 show the feature spaces used in the male and female multistage classifier to discriminate low from high energy utterances.

IV. ANALYSIS

The superior performance of the female prosody classifier seems to be largely an artifact of this approach not scaling well, given the results for three male speakers. In fact, classification approached 83 percent accuracy for a single male speaker. However, it is interesting that the move to a multistage classifier is effective for female voices and not strikingly so for male. The drastic classification improvement in the female data set between the single-stage and multistage classifiers, far more exaggerated than in the male data, suggests that female voice data may be more susceptible to being broken into categories in this fashion. The difference between the two seems to lie chiefly in the relative ease of separating low- and high-energy samples in the female classifier.

The classification problems in male and female data do not seem to be tremendously dissimilar. Approval versus attention, which was the most difficult case in the Kismet work [1], has the lowest accuracy in the male multistage classifier. Female approval versus attention has the highest accuracy of the female multistage classifier, which may be an artifact of the small size of the female data set or may be more significant.

It is interesting that F24, the average of pitch to energy ratio, is as prevalent as pitch mean and more so than energy mean. The degree to which F24 was a useful feature is not entirely obvious from the data as presented here. It is not entirely clear whether this feature was more useful than the pitch and energy means on their own more because it encoded temporal information or because the division operation provided novel information that could not easily be extracted from the two parent features by a learning algorithm. This could be tested be comparing the performance of F24 to a feature which was simply the ratio of F1 to F8, but even lacking that test the prevalence of F1 and F8 over F20, the average of the product of pitch and

High-energy versus low-energy subclassifier space, female data set

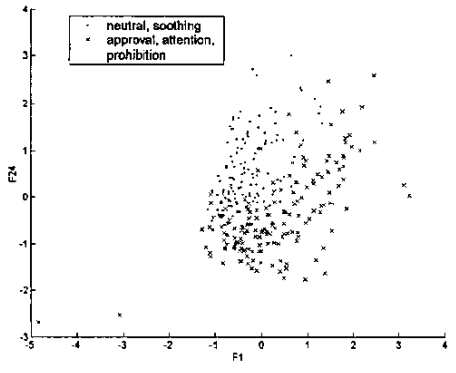

Figure 3. Female data set, high versus low energy classes, features F1 and F24

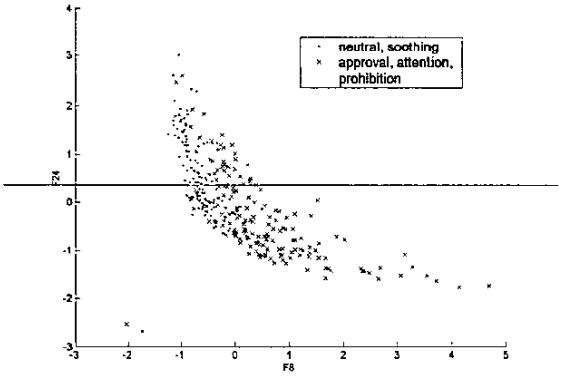

Figure 4. Female data set, high versus low energy classes, features F8 and F24

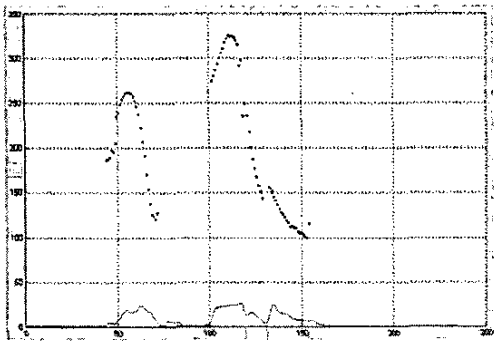

Figure 5. Example of an approval contour in a male voice sample matching Fernald's prototypical contour. The dotted line is pitch as measured on the vertical axis. The solid line is the energy contour scaled to show up on the graph.

energy, which also encodes temporal information, suggests that the division operation may be more important. If the temporal information were more important, it would be interesting to apply some of the techniques used in speech recognition and synthesis to break the samples into phones, perform perhaps ToBI (Tones and Breaks Indices) [8] labeling on those phones, and produce features based on the resulting time sequence. The Festival Speech Synthesis System [9], which is built around the Edinburgh Speech Tools library, provides routines that can perform this kind of intonation labeling. This research direction is related to the BabyEars work [6], which attempted to exploit knowledge about position within the speech sample.

The female prosody classifier clearly performs more poorly than that in the Kismet study [1]. It is not clear whether this is due to coarse data acquisition, differences in feature extraction, the comparatively smaller data set, or a combination of the three. It became evident during data collection that subjects, especially male speakers, had difficulty exhibiting prosody when interacting only with a computer screen, or even a doll.

One somewhat surprising hindrance in classifying male prosody was the difficulty in discriminating between soothing and prohibitory affects, which accounts for at least some of the difference in accuracy between male and female high versus low energy class discrimination. Prohibition and soothing both tend to fall into lower pitch brackets and to have similarly narrow pitch ranges [4, 10], but are expected to differ in their energy statistics.

The example of the subject (one of the experimenters) whose voice admits a high correct classification rate indicates that subjects can train themselves to be more understandable to robots. This is only desirable to the degree that the communication feels reasonably natural for the human. This type of adaptation can be likened to the effort made when speaking to a non-fluent speaker of one's native language, or to someone with a hearing deficiency, or, of course, to an infant. To that degree, effort in speech is precedented, and an acceptable burden to place on the human interactor.

V. Discussion

The key result from this experiment is the observation that there is no critical difference between male and female voice data that makes male prosodic utterances innately much less suitable for classification than those from female subjects. Thus it should be profitable to continue working towards a general classifier that can effectively deal with samples from both genders.

In the next phase of this study, the system has been implemented on the QNX real-time operating system. Audio data is collected continuously from the microphone, and the system begins "paying attention" when it detects segments containing potential speech. Once a sample has been recorded, pitch and energy contours are again extracted using the Edinburgh Speech Tools functions and the results are passed to a C++ implementation of the preprocessor, feature extractor, and recognizer. The process takes .553 seconds from the end of recording on average when the database contains 424 samples. According to Breazeal [1], humans can tolerate interaction delays of up to half a second, so this needs some improvement. This classifier has not yet been rigorously tested but works well on a single speaker, though multi-speaker tests using the first speaker's data for training have so far been disappointing.

Most papers on prosodic classification in this area [1,6] label speech samples by having adult observers listen to and rate them for type and strength. If the goal were to emulate an infant's recognition of prosody directed towards it, it would make sense to classify prosody based explicitly on an infant's reaction to it, if practical. This would allow for better sample fidelity and eliminate the possibility of contamination by higher levels of processing and understanding on the part of the adult observers. A potential experimental setup would be to record video of parent-infant interactions and then label samples based on a combination of the infant's reaction and the observer's perception of the prosody.

Another potential avenue for future research is that of learning a more general model of prosody, rather than just classification, based on ToBI [8] labeling of individual phones[1]. Such a model could then be used for both recognition and synthesis of prosody.

Acknowledgment

The authors would like to thank Frederick Shic and Manfred Lau for their help on this project. Support for this work was provided by a National Science Foundation CAREER award (#0238334). Some parts of the architecture used in this work were constructed under NSF grants #0205542 (ITR: A Framework for Rapid Development of Reliable Robotics Software) and # 0209122 (ITR: Dance, a Programming Language for the Control of Humanoid Robots) and from the DARPA CALO project (BAA 02-21).

[1] We would like to thank Richard Whitney of USC/ISI and Dan Jurafsky of Stanford for conversations which led to this idea.

REFERENCES

[1] C. Breazeal and L. Aryanada, "Recognition of affective communicative intent in robot-directed speech," in 'Proceedings of Humanoids 2000.'

[2] Edinburgh Speech Tools Library, http://www.cstr.ed.ac.uk/projects/speech_tools/

[3] Speech Filing System Vs3.30, http://www.phon.ucl.ac.uk/resource/sfs/

[4] A. Fernald, 'Intonation and communicative intent in mothers' speech to infants: Is the melody the message?", in Child Development 60, pp. 1497-1510.

[5] I.H. Witten and E. Frank, Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, CA.

[6] M. Slaney and G. McRoberts, "Baby Ears: A recognition system for affective vocalizations," in 'Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing', Seattle, WA, May 12-15, 1998.

[7] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech understanding: An overview of recent research at SRI," in 'Proceedings of ISCA Workshop on Prosody in Speech Recognition and Understanding,' 13-16, 2001.

[8] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labeling English prosody," in *Proceedings, Second International Conference on Spoken Language Processing* 2, Banff, Canada, pp. 867-70 (1992).

[9] The Festival Speech Synthesis System, http://www.cstr.ed.ac.uk/projects/festival/

[10] M. Papousek, H. Papousek, and M. Bornstein, "The naturalistic vocal environment of young infants: on the significance of homegeneity and variability in parental speech," in T. Field and N. Fox, eds, 'Social perception in infants,' Ablex, Norwood, NJ, pp. 269-297 (1985