Emotive Qualities in Robot Speech

Cynthia Breazeal

Abstract—
This paper explores the expression of emotion in synthesized speech for an anthropomorphic robot. We have adapted several key emotional correlates of human speech to the robot's speech synthesizer to allow the robot to speak in either an angry, calm, disgusted, fearful, happy, sad, or surprised manner. We have evaluated our approach thorough acoustic analysis of the speech patters for each vocal affect and have studied how well human subjects perceive the intended affect.

Keywords— **Human-robot interaction, emotive expression, synthesized speech.**

I. Introduction

There is a growing research and commercial interest in building robots that can interact with people in a life-like and social manner. For robotic applications where the robot and human establish and maintain a long term relationship, such as robotic pets for children or robotic nursemaids for the elderly, communication of affect is important. There have been a number of projects exploring models of emotion for robots or animated life-like characters [1], [2], [3], [4], [5], the recognition of emotive states in people [6], [7], [8], [9], and the expression of affect in facial expression [10], [11], [12] and body movement [13]. This paper explores the expression of emotion in synthesized speech for an anthropomorphic robot (called Kismet) with a highly expressive face. We have adapted several key emotional correlates of human speech to the robot's synthesizer (based on *DECTALK v4.0*) to allow Kismet to speak in either an angry, calm, disgusted, fearful, happy, sad, or surprised manner. We have evaluated our approach thorough acoustic analysis of the speech patters for each vocal affect. We have also studied how well human subjects perceive the intended affect.

It is well-accepted that facial expressions (related to affect) and facial displays (which serve a communication function) are important for verbal communication. Hence, Kismet's vocalizations should convey the affective state of the robot. This provides a person with important affective information as to how to appropriately engage a sociable robot like Kismet. If done properly, Kismet could then use its emotive vocalizations to convey disapproval, frustration, disappointment, attentiveness, or playfulness. This fosters richer and sustained social interaction, and helps to maintain the person's interest. For a compelling verbal exchange, it is also important for Kismet to accompany its expressive speech with appropriate motor movements of the lips, jaw, and face. The ability to lip synchronize with expressive speech strengthens the perception of Kismet as a social creature that expresses itself vocally and through facial expression. A disembodied voice would be a detriment to a life-like quality of interaction that we would like Kismet to have with people. Synchronized movements of the face with voice both complement as well as supplement the information transmitted through the verbal channel. In earlier work we have presented Kismet's emotion system and its expressive facial animation system (that includes emotive facial expressions and lip synchronization) [12]. This paper presents our work in giving Kismet's voice emotive qualities.

C. Breazeal is with the Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA USA. E-mail: cynthia@ai.mit.edu.

II. Emotion in Speech

There has been an increasing amount of work in identifying those acoustic features that vary with a speaker's affective state [14]. Figure 1 summarizes the effects of emotion in human speech that tend to alter the pitch, timing, voice quality, and articulation of the speech signal [15]. Several of these features, however, are also modulated by the prosodic effects that the speaker uses to communicate grammatical structure and lexical correlates. These tend to have a more localized influence on the speech signal, such as emphasizing a particular word. For recognition tasks, this increases the challenge of isolating those feature characteristics modulated by emotion. Even humans are not perfect at perceiving the intended emotion for those emotional states that have similar acoustic characteristics. For instance, surprise can be perceived or understood as either joyous surprise (happiness) or apprehensive surprise (fear). Disgust is a form of disapproval and can be confused with anger. Picard (1997) [6] presents a nice overview of work in this area.

The effect of emotions on the human voice

	fear	anger	sorrow	joy	disgust	surprise
speech rate	much faster	slightly faster	slightly slower	faster or slower	very much slower	much faster
pitch average	very much higher	very much higher	slightly lower	much higher	very much lower	much higher
pitch range	much wider	much wider	slightly narrower	much wider	slightly wider	
intensity	normal	higher	lower	higher	lower	higher
voice quality	irregular voicing	breathy chest tone	resonant	breathy blaring	grumbled chest tone	
pitch changes	normal	abrupt on stressed syllable	downward inflections	smooth upward inflections	wide downward terminal inflections	rising contour
articulation	precise	tense	slurring	normal	normal	

Fig. 1. Typical effect of emotions on adult human speech, adapted from Murray and Arnott (1993) and Picard (1997).

There have been a few systems developed to synthesize emotional speech. For instance, Jun Sato (see `www.ee.seikei.ac.jp/user/junsato/research/`) trained a neural network to modulate a neutrally spoken speech signal (in Japanese) to convey one of four emotional states (happiness, anger, sorrow, disgust). The neural network was trained on speech spoken by Japanese actors. This approach has the advantage that the output speech signal sounds more natural than purely synthesized speech. For our interactive robot application, this approach has the disadvantage that the speech input to the system must be prerecorded. Kismet must be able to generate its own utterances to suit the circumstance.

The *Affect Editor* by Janet Cahn is among the earliest work in expressive synthesized speech [15]. Her system was based on *DECtalk3*, a commercially available text-to-speech speech synthesizer. Given an English sentence and an emotional quality (one of anger, disgust, fear, joy, sorrow, or surprise), she developed a methodology for mapping the emotional correlates of speech (changes in pitch, timing, voice quality, and articulation) onto the underlying DECtalk synthesizer settings. She took great care to introduce the global prosodic effects of emotion while still preserving the more local influences of grammatical and lexical correlates of speech intonation. With respect to giving Kismet the ability to generate emotive vocalizations, Cahn's work is a valuable resource that we have adapted and extended to suit our purposes.

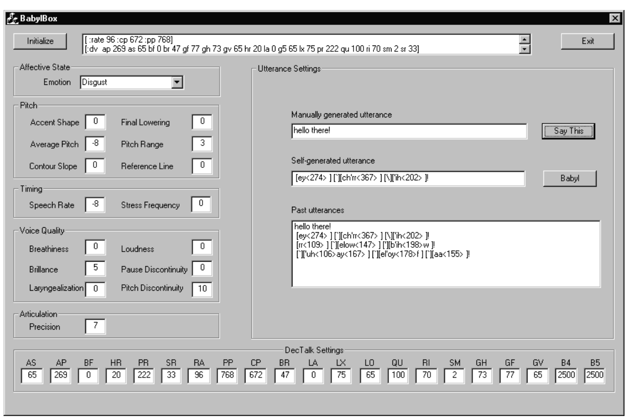

Fig. 2. Kismet's expressive speech GUI. Listed is a selection of emotive qualities, the vocal affect parameters, and the synthesizer settings.

III. The Expressive Voice Synthesis System

Emotions have a global impact on speech since they modulate the respiratory system, larynx, vocal tract, muscular system, heart rate, and blood pressure. There are an assortment of *vocal affect parameters (VAP)* that alter the pitch, timing, voice quality, and articulation aspects of the speech signal. The pitch-related parameters affect the pitch contour of the speech signal, which is the primary contributor for affective information. The pitch-related parameters include *accent shape*, *average pitch*, *pitch contour slope*, *final lowering*, *pitch range*, and *pitch reference line*. The timing-related parameters modify the prosody of the vocalization, often being reflected in speech rate and stress placement. The timing-related parameters include *speech rate*, *pauses*, *exaggeration*, and *stress frequency*. The voice-quality parameters include *loudness*, *brilliance*, *breathiness*, *laryngealization*, *pitch discontinuity*, and *pause discontinuity*. The articulation parameter modifies the precision of what is uttered, either being more enunciated or slurred. These vocal affect parameters are described in more detail below.

Our task is to derive a mapping of these physiological vocal affect parameters to the underlying synthesizer settings (we use *DECTALK v4.0*) to convey the emotional qualities of anger, fear, disgust, happiness, sadness, and surprise in Kismet's voice. There is currently a single fixed mapping per emotional quality. Figure 3 along with the equations presented in this paper summarize how the vocal affect parameters are mapped to the DECtalk synthesizer settings. The default values and max/min bounds for these settings are given in Figure 4. Figure 5 summarizes how each emotional quality of voice is mapped onto the VAPs.

DECtalk Synthesizer Setting	DECtalk Symbol	norm	Controlling Vocal Affect Parameter(s)	Percent of Control
average pitch	ap	.51	average pitch	1
assertiveness	as	.65	final lowering contour direction	.8 .2
baseline fall	bf	0	contour direction final lowering	-.5 .5
breathiness	br	.46	breathiness	1
comma pause	:cp	.238	speech rate	-1
gain of frication	gf	.6	precision of articulation	1
gain of aspiration	gh	.933	precision of articulation	1
gain of voicing	gv	.76	loudness precision of articulation	.6 .4
hat rise	hr	.2	reference line	1
laryngealization	la	0	laryngealization	1
loudness	lo	.5	loudness	1
lax breathiness	lx	.75	breathiness	1
period pause	:pp	.666	speech rate	-1
pitch range	pr	.8	pitch range	1
quickness	qu	.5	pitch discontinuity	1
speech rate	:ra	.2	speech rate	1
richness	ri	.4	brillance	1
smoothness	sm	.05	brillance	-1
stress rise	sr	.220	accent shape pitch discontinuity	.8 .2

Fig. 3. Percent contributions of vocal affect parameters to DECtalk synthesizer settings. The absolute values of the contributions in the far right column add up to 1 (100%) for each synthesizer setting. See the equations in section ?? for the mapping.

A. The Vocal Affect Parameters (VAPs)

The following six *pitch parameters* influence the pitch contour of the spoken utterance. The pitch contour is the trajectory of the fundamental frequency, f_0, over time.

• *Accent Shape*: Modifies the shape of the pitch contour for any pitch accented word by varying the rate of f_0 change about that word. A high accent shape corresponds to speaker agitation where there is a high peak f_0 and a steep rising and falling pitch contour slope. This parameter has a substantial contribution to DECtalk's `stress rise` setting, which regulates the f_0 magnitude of pitch-accented words.

• *Average Pitch*: Quantifies how high or low the speaker appears to be speaking relative to their normal speech. It is the average f_0 value of the pitch contour. It varies directly with DECtalk's `average pitch`.

• *Contour Slope*: Describes the general direction of the pitch contour, which can be characterized as rising, falling, or level. It contributes to two DECtalk settings. It has a small contribution to the `assertiveness` setting, and varies inversely with the `baseline fall` setting.

• *Final Lowering*: Refers to the amount that the pitch contour falls at the end of an utterance. In general, an utterance will sound emphatic with a strong final lowering, and tentative if weak. It can also be used as an auditory cue to regulate turn taking. A strong final lowering can signify the end of a speaking turn, whereas a speaker's intention to continue talking can be conveyed with a slight rise at the end. This parameter strongly contributes to DECtalk's `assertiveness` setting and somewhat to the `baseline fall` setting.

• *Pitch Range*: Measures the bandwidth between the maximum and minimum f_0 of the utterance. The pitch range expands and contracts about the average f_0 of the pitch contour. It varies directly with DECtalk's `pitch range` setting.

• *Reference Line*: Controls the reference pitch f_0 contour. Pitch accents cause the pitch trajectory to rise above or dip below this reference value. DECtalk's `hat rise` setting very roughly approximates this.

The vocal affect *timing parameters* contribute to speech rhythm. Such correlates arise in emotional speech from physiological changes in respiration rate (changes in breathing patterns) and level of arousal.

• *Speech Rate*: Controls the rate of words or syllables uttered per minute. It influences how quickly an individual word or syllable is uttered, the duration of sound to silence within an utterance, and the relative duration of phoneme classes. Speech is faster with higher arousal and slower with lower arousal. This parameter varies directly with DECtalk's `speech rate` setting. It varies inversely with DECtalk's `period pause` and `comma pause` settings as faster speech is accompanied with shorter pauses.

• *Stress Frequency*: Controls the frequency of occurrence of pitch accents and determines the smoothness or abruptness of f_0 transitions. As more words are stressed, the speech sounds more emphatic and the speaker more agitated. It filters other vocal affect parameters such as precision of articulation and accent shape, and thereby contributes to the associated DECtalk settings.

Emotion can induce not only changes in pitch and tempo, but in voice quality as well. These phenomena primarily arise from changes in the larynx and articulatory tract. The *voice quality* parameters are as follows:

DECtalk Synthesizer Setting	unit	neutral setting	min setting	max setting
average pitch	Hz	306	260	350
assertiveness	%	65	0	100
baseline fall	Hz	0	0	40
breathiness	dB	47	40	55
comma pause	msec	160	-20	800
gain of frication	dB	72	60	80
gain of aspiration	dB	70	0	75
gain of voicing	dB	55	65	68
hat rise	Hz	20	0	80
laryngealization	%	0	0	10
loudness	dB	65	60	70
lax breathiness	%	75	100	0
period pause	msec	640	-275	800
pitch range	%	210	50	250
quickness	%	50	0	100
speech rate	wpm	180	75	300
richness	%	40	0	100
smoothness	%	5	0	100
stress rise	Hz	22	0	80

Fig. 4. Default DECtalk synthesizer settings for Kismet's voice that are used in the equations for altering these values to produce Kismet's expressive speech.

• *Breathiness*: Controls the aspiration noise in the speech signal. It adds a tentative and weak quality to the voice, when speaker is minimally excited. DECtalk `breathiness` and `lax breathiness` vary directly with this.

• *Brillance*: Controls the perceptual effect of relative energies of the high and low frequencies. When agitated, higher frequencies predominate and the voice is harsh or "brilliant". When speaker is relaxed or depressed, lower frequencies dominate and the voice sounds soothing and warm. DECtalk's `richness` setting varies directly as it enhances the lower frequencies. In contrast, DECtalk's `smoothness` setting varies inversely since it attenuates higher frequencies.

• *Laryngealization*: Controls the perceived creaky voice phenomena. It arises from minimal sub-glottal pressure and a small open quotient such that f_0 is low, the glottal pulse is narrow, and the fundamental period is irregular. It varies directly with DECtalk's `laryngealization` setting.

• *Loudness*: Controls the amplitude of the speech waveform. As a speaker becomes aroused, the sub-glottal pressure builds which increases the signal amplitude. As a result, the voice sounds louder. It varies directly with DECtalk's `loudness` setting. It also influences DECtalk's `gain of voicing`.

• *Pause Discontinuity*: Controls the smoothness of f_0 transitions from sound to silence for unfilled pauses. Longer or more abrupt silences correlate with being more emotionally upset. It varies directly with DECtalk's `quickness` setting.

• *Pitch Discontinuity*: Controls smoothness or abruptness of f_0 transitions, and the degree to which the intended targets are reached. With more speaker control, the transitions are smoother. With less control, they transitions are more abrupt. It contributes to DECtalk's `stress rise` and `quickness` settings.

The autonomic nervous system modulates articulation by inducing an assortment of physiological changes such as causing dryness of mouth or increased salivation. There is only one *articulation parameter* as follows:

• *Precision*: Controls a range of articulation from enunciation to slurring. Slurring has minimal frication noise, whereas greater enunciation for consonants results in increased frication. Stronger enunciation also results in an increase in aspiration noise and voicing. The precision of articulation varies directly with DECtalk's `gain of frication`, `gain of voicing`, and `gain of aspiration`.

	Anger	disgust	fear	happy	sad	surprise	neutral
accent shape	10	0	10	10	-7	9	0
average pitch	-10	-10	10	3	-7	6	0
contour slope	10	0	10	0	0	10	0
final lowering	10	5	-10	-4	8	-10	0
pitch range	10	5	10	10	-10	10	0
reference line	-10	0	10	-8	-1	-8	0
speech rate	4	-8	10	3	-6	6	0
stress frequency	0	0	10	5	1	0	0
breathiness	-5	0	0	-5	0	-9	0
brillance	10	5	10	-2	-6	9	0
laryngealization	0	0	-10	0	0	0	0
loudness	10	-5	10	8	-5	10	0
pause discontinuity	10	0	10	-10	-8	-10	0
pitch discontinuity	3	10	10	6	0	10	0
precision of articulation	10	7	0	-3	-5	0	0

Fig. 5. The mapping from each expressive quality of speech to the vocal affect parameters (VAPs). There is a single fixed mapping for each emotional quality.

B. Mapping VAPs to Synthesizer Settings

This section presents the equations that map the vocal affect parameters to synthesizer setting values. Linear changes in these vocal affect parameter values result in a non-linear change in the underlying synthesizer settings. Furthermore, the mapping between parameters and synthesizer settings is not necessarily one-to-one. Each parameter affects a percent of the final synthesizer setting's value (figure 3). When a synthesizer setting is modulated by more than one parameter, its final value is the sum of the effects of the controlling parameters. The total of the absolute values of these percentages must be 100%. See figure 4 for the allowable bounds of synthesizer settings. The computational mapping occurs in three stages. The vocal affect parameters can assume integer values within the range of $(-10, 10)$. Negative numbers correspond to lesser effects, positive numbers correspond to greater effects, and zero is the neutral setting. These values are set according to the current specified emotion as shown in figure 5.

In the first stage, the percentage of each of the VAPs (VAP_i) to its total range is computed, (PP_i). This is given by the equation:

$$PP_i = \frac{VAP_{value_i} + VAP_{offset}}{VAP_{max} - VAP_{min}}$$

VAP_i is the current VAP under consideration, VAP_{value} is its value specified by the current emotion, $VAP_{offset} = 10$ adjusts these values to be positive, $VAP_{max} = 10$, and $VAP_{min} = -10$.

In the second stage, a weighted contribution ($WC_{j,i}$) of those VAP_i that control each of DECtalk's synthesizer settings (SS_j) is computed. The far right column of figure 3 specifies each of the corresponding *scale factors* ($SF_{j,i}$). Each scale factor represents a percentage of control that each VAP_i applies to its synthesizer setting SS_j.

For each synthesizer setting, SS_j:
  For each corresponding scale factor, $SF_{j,i}$ of VAP_i:
    If $SF_{j,i} \geq 0$
      $WC_{j,i} = PP_i \times SF_{j,i}$
    If $SF_{j,i} \leq 0$
      $WC_{j,i} = (1 - PP_i) \times (-SF_{j,i})$
  $SS_j = \sum_i WC_{j,i}$

At this point, each synthesizer value has a value $0 \leq SS_j \leq 1$. In the final stage, each synthesizer setting SS_j is scaled about 0.5. This produces the final synthesizer value, $SS_{j_{final}}$. The final value is sent to the speech synthesizer. The maximum, Minimum, and default values of the synthesizer settings are shown in figure 4.

For each final synthesizer setting, $SS_{j_{final}}$:
  Compute $SS_{j_{offset}} = SS_j - norm$
  If $SS_{j_{offset}} \geq 0$
    $SS_{j_{final}} = SS_{j_{default}} +$
      $(2 \times SS_{j_{offset}} \times (SS_{j_{max}} - SS_{j_{min}}))$
  If $SS_{j_{offset}} \leq 0$
    $SS_{j_{final}} = SS_{j_{default}} +$
      $(2 \times SS_{j_{offset}} \times (SS_{j_{default}} - SS_{j_{min}}))$

IV. Kismet's Expressive Utterances

Given a string to be spoken and the updated synthesizer settings, Kismet can vocally express itself with different emotional qualities (anger, disgust, fear, joy, sorrow, or surprise). To evaluate Kismet's speech, we analyzed the produced utterances with respect to the acoustical correlates of emotion. This reveals whether the implementation produces similar acoustical changes to the speech waveform given a specified emotional state. We also evaluated how the affective modulations of the synthesized speech are perceived by human listeners.

.1 Analysis of Speech

To analyze the performance of the expressive vocalization system, we extracted the dominant acoustic features that are highly correlated with emotive state. The acoustic features and their modulation with emotion are summarized in figure 1. Specifically, these are average pitch,

pitch range, pitch variance, and mean energy. To measure speech rate, we extracted the overall time to speak and the total time of voiced segments.

	nzpmean	nzpvar	pmax	pmin	prange	egmean	length	voiced	unvoiced
anger-city	292.5	6348.7	444.4	166.7	277.7	112.2	81	52	29
anger-moved	269.1	4703.8	444.4	160	284.4	109.8	121	91	30
anger-picture	273.2	6850.3	444.4	153.8	290.6	110.2	112	51	61
anger-average	278.3	5967.6	444.4	160.17	284.2	110.7	104.6	64.6	40
calm-city	316.8	802.9	363.6	250	113.6	102.6	85	58	27
calm-moved	304.5	897.3	363.6	266.7	96.9	103.6	124	94	30
calm-picture	302.2	1395.5	363.6	235.3	128.3	102.4	118	73	45
calm-average	307.9	1031.9	363.6	250.67	112.93	102.9	109	75	34
disgust-city	268.4	2220.0	400	173.9	226.1	102.5	124	83	41
disgust-moved	264.6	1669.2	400	190.5	209.5	101.6	173	123	50
disgust-picture	275.2	3264.1	400	137.9	262.1	102.3	157	82	75
disgust-average	269.4	2384.4	400	167.4	232.5	102.1	151.3	96	55.3
fear-city	417.0	8986.7	500	235.3	264.7	102.8	59	27	32
fear-moved	357.2	7145.5	500	160	340	102.6	89	53	36
fear-picture	388.2	8830.9	500	160	340	103.6	86	41	45
fear-average	387.4	8321.0	500	185.1	314.9	103.0	78	40.3	37.6
happy-city	388.3	5810.6	500	285.7	214.3	106.6	71	54	17
happy-moved	348.2	6188.8	500	173.9	326.1	109.2	109	78	31
happy-picture	357.7	6038.3	500	266.7	233.3	106.0	100	57	43
happy-average	364.7	6012.6	500	242.1	257.9	107.2	93.3	63	30.3
sad-city	279.8	77.9	285.7	266.7	19	98.6	88	62	26
sad-moved	276.9	90.7	285.7	266.7	19	99.1	144	93	51
sad-picture	275.5	127.2	285.7	250	35.7	98.3	138	83	55
sad-average	277.4	98.6	285.7	261.1	24.5	98.7	123.3	79.3	44
surprise-city	394.3	8219.4	500	148.1	351.9	107.5	69	49	20
surprise-moved	360.3	7156.0	500	160	340	107.8	101	84	17
surprise-picture	371.6	8355.7	500	285.7	214.3	106.7	98	54	44
surprise-average	375.4	7910.4	500	197.9	302.0	107.3	89.3	62.3	27

Fig. 6. Table of acoustic features for the three utterances.

Features were extracted from three phrases:

- *Look at that picture*
- *Go to the city*
- *It's been moved already*

The results are summarized in figure 6. The values for each feature are displayed for each phrase with each emotive quality (including the neutral state). The averages are also presented in the table and plotted in figure 7. These plots easily illustrate the relationship of how each emotive quality modulates these acoustic features with respect to one another. The pitch contours for each emotive quality are shown in figure 8. They correspond to the utterance "It's been moved already." Relating these plots with figure 1, it is clear that many of the acoustic correlates of emotive speech are preserved in Kismet's speech.

Kismet's vocal quality varies with its emotive state as follows:

- *Fearful speech* is very fast with wide pitch contour, large pitch variance, very high mean pitch, and normal intensity. I have added a slightly breathy quality to the voice as people seem to associate it with a sense of trepidation.
- *Angry speech* is loud and slightly fast with a wide pitch range and high variance. We've purposefully implemented a low mean pitch to give the voice a prohibiting quality. This differs from figure 1, but a preliminary study demonstrated a dramatic improvement in recognition performance of naive subjects. This makes sense as it gives the voice a threatening quality.
- *Sad speech* has a slower speech rate, with longer pauses than normal. It has a low mean pitch, a narrow pitch range and low variance. It is softly spoken with a slight breathy quality. This differs from figure 1, but it gives the voice a tired quality. It has a pitch contour that falls at the end.

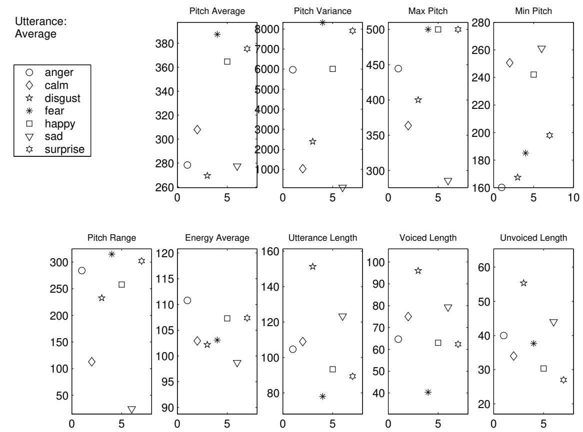

Fig. 7. Plots of acoustic features of Kismet's speech. Plots illustrate how each emotion relates to the others for each acoustic feature. The horizontal axis simply maps an integer value to each emotion for ease of viewing (anger=1, calm=2, etc.)

- *Happy speech* is relatively fast, with a high mean pitch, wide pitch range, and wide pitch variance. It is loud with smooth undulating inflections as shown in figure 8.
- *Disgusted speech* is slow with long pauses interspersed. It has a low mean pitch with a slightly wide pitch range. It is fairly quiet with a slight creaky quality to the voice. The contour has a global downward slope as shown in figure 8.
- *Surprised speech* is fast with a high mean pitch and wide pitch range. It is fairly loud with a steep rising contour on the stressed syllable of the final word.

.2 Human Listener Experiments

To evaluate Kismet's expressive speech, nine subjects were asked to listen to prerecorded utterances and to fill out a forced-choice questionnaire. Subjects ranged from 23 to 54 years of age, all affiliated with MIT. The subjects had very limited to no familiarity with Kismet's voice.

In this study, each subject first listened to an introduction spoken with Kismet's neutral expression. This was to acquaint the subject with Kismet's synthesized quality of voice and neutral affect. A series of eighteen utterances followed, covering six expressive qualities (anger, fear, disgust, happiness, surprise, and sorrow). Within the experiment, the emotive qualities were distributed randomly. Given the small number of subjects per study, we only used a single presentation order per experiment. Each subject could work at his/her own pace and control the number of presentations of each stimulus.

The three stimulus phrases were: "I'm going to the city," "I saw your name in the paper," and "It's happening tomorrow." The first two test phrases were selected because Cahn (1990) had found the word choice to have reasonably neutral affect. In a previous version of the study, subjects reported that it was just as easy to map emotional correlates onto English phrases as to Kismet's non-linguistic vocalizations (akin to infant-like babbles). Their performance for English phrases and Kismet's babbles supports this.

Using a forced choice paradigm, the subjects were sim-

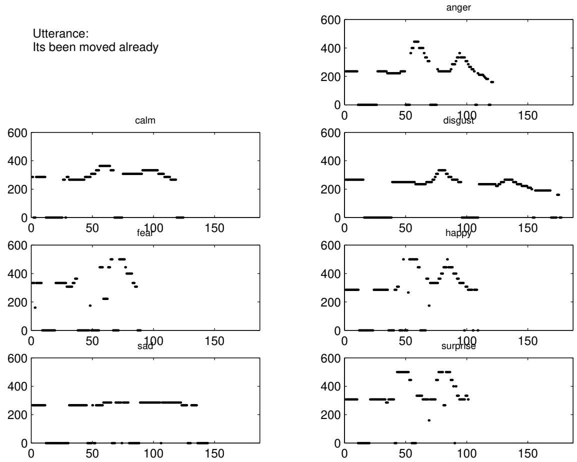

Fig. 8. Pitch analysis of Kismet's speech for the English phrase "It's been moved already."

forced choice percentage (random=17%)

	anger	disgust	fear	happy	sad	surprise	% correct
anger	75	15	0	0	0	10	75/100
disgust	21	50	4	0	25	0	50/100
fear	4	0	25	8	0	63	25/100
happy	0	4	4	67	8	17	67/100
sad	8	8	0	0	84	0	84/100
surprise	4	0	25	8	4	59	59/100

Fig. 9. Naive subjects assessed the emotion conveyed in Kismet's voice in a forced-choice evaluation. All emotional qualities were recognized with reasonable performance except for "fear" which was most often confused for "surprise/excitement." Both expressive qualities share high arousal, so the confusion is not unexpected.

ply asked to circle the word which best described the voice quality. The choices were "anger," "disgust," "fear/panic," "happy," "sad," "surprise/excited." From a previous iteration of the study, we found that word choice mattered. A given emotion category can have a wide range of vocal affects. For instance, the subject could interpret "fear" to imply "apprehensive," which might be associated with Kismet's whispery vocal expression for sadness. Alternatively, it could be associated with "panic" which is a more aroused interpretation. The results from these evaluations are summarized in figure 9.

Overall, the subjects exhibited reasonable performance in correctly mapping Kismet's expressive quality with the targeted emotion. However, the expression of "fear" proved somewhat problematic. For all other expressive qualities, the performance was significantly above random. Furthermore, misclassifications were highly correlated to similar emotions. For instance, "anger" was sometimes confused with "disgust" (sharing negative valence) or "surprise/excitement" (both sharing high arousal). "Disgust" was confused with other negative emotions. "Fear" was confused with other high arousal emotions (with "surprise/excitement" in particular). The distribution for "happy" was more spread out, but it was most often confused with "surprise/excitement," with which it shares high arousal. Kismet's "sad" speech was confused with other negative emotions. The distribution for "surprise/excitement" was broad, but it was most often confused for "fear."

V. Summary

For the purposes of evaluation, the current set of data is promising. Misclassifications are particularly informative. The mistakes are highly correlated with similar emotions, which suggests that arousal and valence are conveyed to people (arousal being more consistently conveyed than valence). We are using the results of this study to improve Kismet's expressive qualities. In addition, Kismet expresses itself through multiple modalities, not just through voice. We believe that Kismet's facial expression and body posture should help resolve the ambiguities encountered through voice alone.

Acknowledgments

This work was supported in part by DARPA under contract DABT 63-99-1-0012 and in part by NTT.

References

[1] Juan Velasquez, "Modeling emotions and other motivations in synthetic agents," in *Proceedings of the 1997 National Conference on Artificial Intelligence, AAAI97*, july 1997, pp. 10–15.

[2] C. Breazeal, "Robot in society: friend or appliance?," in *Proceedings of Agents99 workshop on Emotion-based architectures*, Seattle, WA, 1999, pp. 18–26.

[3] D. Canamero, "Modeling motivations and emotions as a basis for intelligent behavior," in *Proceedings of the First International Conference on Autonomous Agents*, L. Johnson, Ed. 1997, pp. 148–155, ACM Press.

[4] S.Y. Yoon, B. Blumberg, and G. Schneider, "Motivation driven learning for interactive synthetic characters," in *Proceedings of Agents2000 (to appear)*, 2000.

[5] S. Reilly, *Believable Social and Emotional Agents*, Ph.D. thesis, CMU School of Computer Science, Pittsburgh, PA, 1996.

[6] R. Picard, *Affective Computation*, MIT Press, Cambridge, MA, 1997.

[7] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," in *Proceedings of Humanoids2000 (submitted)*, 2000.

[8] D. Roy and A. Pentland, "Automatic spoken affect analysis and classification," *Proceedings of the 1996 international conference on automatic face and gesture recognition*, 1996.

[9] L. Chen and T. Huang, "Multimodal human emotion/expression recognition," *Proceedings of the second international conference on automatic face and gesture recognition*, pp. 366–371, April 1998.

[10] F. Hara, "Personality characterization of animate face robot through interactive communication with human," in *Proceedings of IARP98*, Tsukuba, Japan, 1998, pp. IV–1.

[11] H. Takanobu, A. Takanishi, S. Hirano, I. Kato, K. Sato, and T. Umetsu, "Development of humanoid robot heads for natural human-robot communication," in *Proceedings of HURO98*, 1998.

[12] C. Breazeal, *Socialbe Machines: Expressive Social Exchange Between Humans and Robots*, Ph.D. thesis, Massachusetts Institute of Technology, department of Electrical Engineering and Computer Science, Cambridge, MA, May 2000.

[13] C. Rose, B. Bodenheimer, and M. Cohen, “Verbs and adverbs,” *In press: Computer Graphics and Animation*, 1998.
[14] I. Murray and L. Arnott, “Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion,” *Journal Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.
[15] J. Cahn, “Generating expression in synthesized speech,” M.S. thesis, MIT Media Lab, Cambridge, MA, 1990.