Advanced Robotics, Vol. 17, No. 2, pp. 97–113 (2003)

Also available online - www.vsppub.com

Full paper

Emotive qualities in lip-synchronized robot speech

CYNTHIA BREAZEAL *
MIT Media Lab, 77 Massachusetts Ave NE18-5fl, Cambridge, MA, USA

Received 11 April 2002; accepted 31 May 2002

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 patterns for each vocal affect and have studied how well human subjects perceive the intended affect. The robot lip synchronizes in real-time to enhance the delivery of its expressive utterances.

Keywords: Human–robot interaction; emotive expression; synthesized speech; real-time lip synchronization.

1. 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, the recognition of emotive states in people, and the expression of affect in facial expression and body movement [1–5]. 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.

*E-mail: cynthiab@media.mit.edu

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) [6]. This paper presents our work in giving *Kismet*'s voice emotive qualities.

2. EMOTION IN SPEECH

There has been an increasing amount of work in identifying those acoustic features that vary with a speaker's affective state [7]. Table 1 summarizes the effects of emotion in human speech that tend to alter the pitch, timing, voice quality and articulation of the speech signal [8]. 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 (i.e. happiness) or apprehensive surprise (i.e. fear). Disgust is a form of disapproval and can be confused with anger. Picard [1] presents a nice overview of work in this area.

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 (i.e. happiness, anger, sorrow or 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 [8]. Her system was based on *DECtalk3*, a commercially

Table 1.
Typical effect of emotions on adult human speech, adapted from [1, 7]

	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 narrower	
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	

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 including real-time lip synchronization with complementary facial expression and models of emotion.

3. EXPRESSIVE VOICE SYNTHESIS

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* (VAPs) that alter the pitch, timing, voice quality and articulation aspects of the speech signal (summarized in Cahn [8]). 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

Table 2.
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)

DECtalk synthesizer setting	DECtalk symbol	Controlling vocal affect parameter(s)	Percent of control
Average pitch	ap	average pitch	1
Assertiveness	as	final lowering	0.8
		contour direction	0.2
Baseline fall	bf	contour direction	−0.5
		final lowering	0.5
Breathiness	br	breathiness	1
Comma pause	:cp	speech rate	−1
Gain of frication	gf	precision of articulation	1
Gain of aspiration	gh	precision of articulation	1
Gain of voicing	gv	loudness	1
		precision of articulation	0.6 0.4
Hat rise	hr	reference line	1
Laryngealization	la	laryngealization	1
Loudness	lo	loudness	1
Lax breathiness	lx	breathiness	1
Period pause	:pp	speech rate	−1
Pitch range	pr	pitch range	1
Quickness	qu	pitch discontinuity	1
Speech rate	:ra	speech rate	1
Richness	ri	brillance	1
Smoothness	sm	brillance	−1
Stress rise	sr	accent shape	0.8
		pitch discontinuity	0.2

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. Table 2 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

Table 3.
Default DECtalk synthesizer settings for *Kismet*'s voice that are used in the equations for altering these values to produce *Kismet*'s expressive speech

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	ms	160	−20	800
Gain of frication	dB	72	60	80
Gain of aspiration	dB	70	0	75
Gain of voicing	dB	65	55	68
Hat rise	Hz	20	0	80
Laryngealization	%	0	0	10
Loudness	dB	65	60	70
Lax breathiness	%	75	0	100
Period pause	ms	640	−275	800
Pitch range	%	210	50	250
Quickness	%	50	0	100
Speech rate	w.p.m.	180	75	300
Richness	%	40	0	100
Smoothness	%	5	0	100
Stress rise	Hz	22	0	80

bounds for these settings are given in Table 3. Table 4 summarizes how each emotional quality of voice is mapped onto the VAPs.

3.1. VAPs

Below we give a brief description of the VAPs as identified by Cahn. 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.
- *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.
- *Contour slope*. Describes the general direction of the pitch contour, which can be characterized as rising, falling or level.
- *Final lowering*. Refers to the amount that the pitch contour falls at the end of an utterance.
- *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.

Table 4.
The mapping from each expressive quality of speech to the VAPs (there is a single fixed mapping for each emotional quality)

VAP	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 discontinnity	10	0	10	−10	−8	−10	0
Pitch discontinnity	3	10	10	6	0	10	0
Precision of articulation	10	7	0	−3	−5	0	0

- *Reference line*. Controls the reference pitch f_0 contour. Pitch accents cause the pitch trajectory to rise above or dip below this reference value.

The vocal affect *timing parameters* contribute to speech rhythm. Such correlates arise in emotional speech from physiological changes in respiration rate (i.e. 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.
- *Stress frequency*. Controls the frequency of occurrence of pitch accents and determines the smoothness or abruptness of f_0 transitions.

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:

- *Breathiness*. Controls the aspiration noise in the speech signal.
- *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.
- *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.
- *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.
- *Pause discontinuity*. Controls the smoothness of f_0 transitions from sound to silence for unfilled pauses.
- *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.

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.

3.2. VAP to synthesizer settings

This section presents the equations that map the VAPs to synthesizer setting values. Using Cahn's work as a guide, we had to modify her specified parameter values and equations to suit our purposes with *Kismet*. Linear changes in these VAP 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 (Table 2). 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 Table 3 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 Table 4.

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:

$$PP_i = \frac{VAP_{\text{value\_}i} + VAP_{\text{offset}}}{VAP_{\max} - VAP_{\min}},$$

VAP_i is the current VAP under consideration, $VAP_{\text{value\_}i}$ is its value specified by the current emotion, $VAP_{\text{offset}} = 10$ 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 Table 2 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 $SF_{j,i} \geqslant 0$
$WC_{j,i} = PP_i \times SF_{j,i}$
If $SF_{j,i} < 0$
$WC_{j,i} = (1 - PP_i) \times (-SF_{j,i})$
$SS_j = \sum_i WC_{j,i}$.

At this point, each synthesizer value has a value $0 \leqslant SS_j \leqslant 1$. 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 Table 3.

For each final synthesizer setting, SS_{j_final}:

Compute $SS_{j\_\text{offset}} = SS_j - 0.5$
If $SS_{j\_\text{offset}} \geqslant 0$
$SS_{j\_\text{final}} = SS_{j\_\text{default}} +$
$(2 \times SS_{j\_\text{offset}} \times (SS_{j\_\max} - SS_{j\_\text{default}}))$
If $SS_{j\_\text{offset}} < 0$
$SS_{j\_\text{final}} = SS_{j\_\text{default}} +$
$(2 \times SS_{j\_\text{offset}} \times (SS_{j\_\text{default}} - SS_{j\_\min}))$.

4. EXPRESSIVE UTTERANCES

Given a string to be spoken and the updated synthesizer settings, *Kismet* can vocally express itself with different emotional qualities (i.e. 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.

4.1. 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 Table 1. To measure speech rate, we extracted the overall time to speak and the total time of voiced segments. The averages of the values for pitch average, pitch variance, max pitch, min pitch, pitch range, energy average, utterance length, voiced length and unvoiced length for the three test phrases (listed below) are plotted in Fig. 1. These plots easily illustrate the relationship of how each emotive quality modulates these acoustic features with respect to one another.

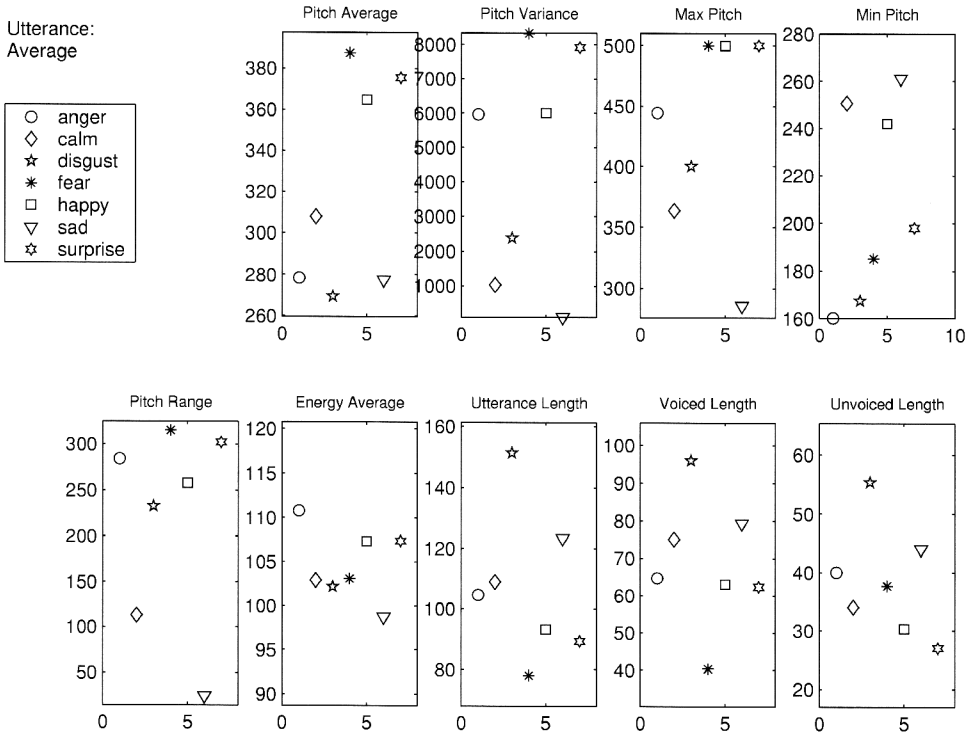

Figure 1. Plots of acoustic features of *Kismet*'s speech. Plots illstrate how each emotion relates to the others for each acoustic feature. The horizontal axis simply maps an integer value to each emotion for ease of viewing (anger = 1, calm = 2, etc.)

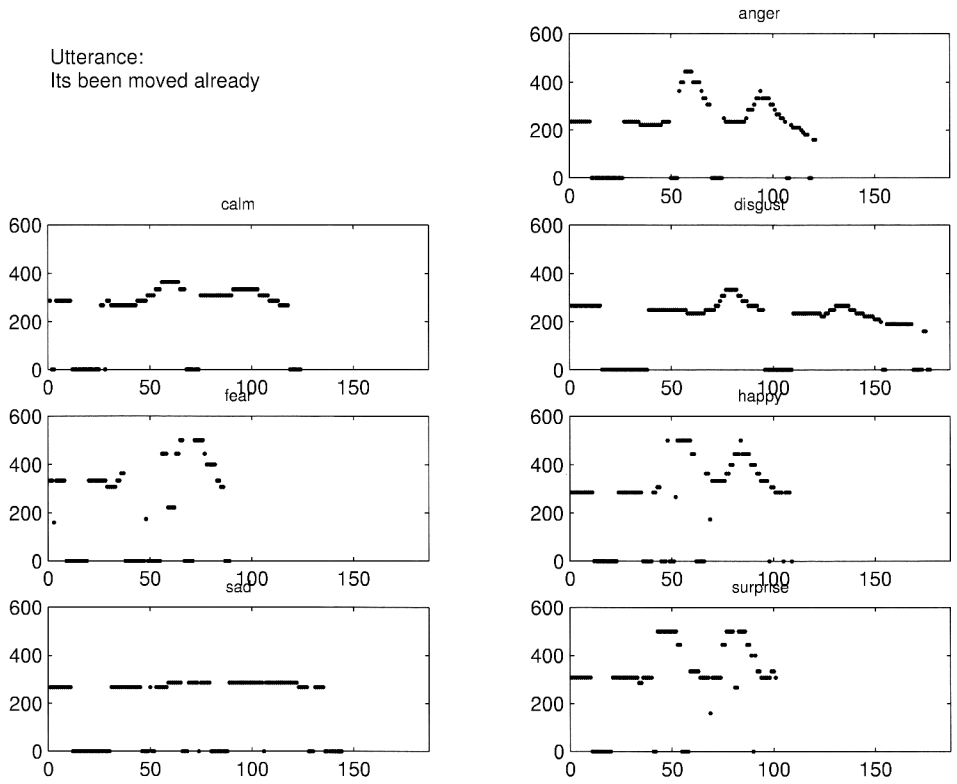

Figure 2. Pitch analysis of *Kismet*'s speech for the English phrase 'It's been moved already'.

Features were extracted from three phrases:

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

A set of pitch contours for the phrase 'Its been moved already' with different emotive qualities is shown in Fig. 2. As illustrated by this sample set, *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 have purposefully implemented a low mean pitch to give the voice a prohibiting quality. This differs from Fig. 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 Fig. 1, but it gives the voice a tired quality. It has a pitch contour that falls at the end.
- *Happy speech* is relatively fast, with a high mean pitch, wide pitch range and wide pitch variance. It is loud with smooth undulating inflections.
- *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.
- *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.

4.2. 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 18 utterances followed, covering six expressive qualities (i.e. 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 [8] 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 simply asked to circle the word which best described the voice quality. The choices were 'anger', 'disgust', 'fear/panic', 'happy', 'sad' and '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 Fig. 3.

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

Figure 3. 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.

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.'

5. REAL-TIME LIP SYNCHRONIZATION AND FACIAL ANIMATION

Given *Kismet*'s ability to express itself vocally, it is important that the robot also be able to support this vocal channel with coordinated facial animation. This includes synchronized lip movements to accompany speech along with facial animation to lend additional emphasis to the stressed syllables. These complementary motor modalities greatly enhance the robot's delivery when it speaks, giving the

impression that the robot 'means' what it says. This makes the interaction more engaging for the human and facilitates proto-dialogue.

5.1. Guidelines from animation

The earliest examples of lip synchronization for animated characters dates back to the 1940s in classical animation [12] and back to the 1970s for computer-animated characters [13]. In these early works, all of the lip animation was crafted by hand (a very time-consuming process). Over time, a set of guidelines evolved that are largely adhered to by animation artists today [14].

According to Madsen, 'simplicity is the secret to successful lip animation'. Extreme accuracy for cartoon animation often looks forced or unnatural. Thus, the goal in animation is not to always imitate realistic lip motions, 'but to create a visual shorthand that passes unchallenged by the viewer' [14]. As the realism of the character increases, however, the accuracy of the lip synchronization follows.

Kismet is a fanciful and cartoon-like character, so the guidelines for cartoon animation apply. In this case, the guidelines suggest that the animator focus on vowel lip motions (especially *o* and *w*) accented with consonant postures (*m*, *b*, *p*) for lip closing. Precision of these consonants gives credibility to the generalized patterns of vowels. The transitions between vowels and consonants should be reasonable approximations of lip and jaw movement. Fortunately, more latitude is granted for more fanciful characters. The mechanical response time of *Kismet*'s lip and jaw motors places strict constraints on how fast the lips and jaw can transition from posture to posture. Madsen also stresses that care must be taken in conveying emotion, as the expression of voice and face can change dramatically.

5.2. Extracting lip synch info

To implement lip synchronization on *Kismet*, a variety of information must be computed in real-time from the speech signal. By placing DECtalk in *memory mode* and issuing the command string (utterance with synthesizer settings), the DECtalk software generates the speech waveform and writes it to memory (a 11.025 kHz waveform). In addition, DECtalk extracts time-stamped phoneme information. From the speech waveform, one can compute its time-varying energy over a window size of 335 samples, taking care to synchronize the phoneme and energy information, and send (*phoneme(t)*, *energy(t)*) pairs to the QNX machine at 33 Hz to coordinate jaw and lip motor control. A similar technique using DECtalk's phoneme extraction capability is reported by Waters and Levergood [15] for real-time lip synchronization for computer-generated facial animation.

To control the jaw, the QNX machine receives the phoneme and energy information, and updates the commanded jaw position at 10 Hz. The mapping from energy to jaw opening is linear, bounded within a range where the minimum position corresponds to a closed mouth and the maximum position corresponds to an open mouth characteristic of surprise. Using only energy to control jaw position produces a

lively effect, but has its limitations [16]. For *Kismet*, the phoneme information is used to make sure that the jaw is closed when either a *m*, *p* or *b* is spoken, or there is silence. This may not necessarily be the case if only energy were used.

Upon receiving the phoneme and energy information from the vocalization system, the QNX vocal communication process passes this information to the motor skill system via the DPRAM. The motor skill system converts the energy information into a measure of facial emphasis (linearly scaling the energy), which is then passed onto the lip synchronization and facial animation processes of the face control motor system. The motor skill system also maps the phoneme information onto lip postures and passes this information to the *lip synchronization* and *facial animation* processes of the motor system that controls the face.

The computer network involved in lip synchronization is a bit convoluted, but supports real-time performance. Within the NT machine, there is a latency of approximately 250 ms from the time the synthesizer generates the speech signal and extracts phoneme information until that speech signal is sent to the sound card. Immediately following the generation and feature extraction phase, the NT machine sends this information to the QNX node that controls the jaw motor. The latency of this stage is less than 1 ms. Within QNX, the energy signal and phoneme information are used to compute the jaw position. To synchronize jaw movement with sound production from the sound card, the jaw command position is delayed by 250 ms. For the same reason, the QNX machine delays the transfer of energy and phoneme information by 100 ms to the *L*-based machines (*L* is a multi-threated Lisp based language). Dual-ported RAM communication is sub-millisecond. The lip synchronization processes running on *L* polls and updates their energy and phoneme values at 40 Hz, much faster than the phoneme information is changing and much faster than the actuators can respond. Energy is scaled to control the amount of facial emphasis and the phonemes are mapped to lip postures. The lip synchronization performance is well-coordinated with speech output since the delays and latencies are fairly consistent.

Kismet's ability to lip-sync within its limits greatly enhances the perception that it is genuinely talking (instead of being some disembodied speech system). It also contributes to the life-like quality and charm of the robot's behavior.

Figure 4 shows how the 50 DECtalk phonemes are mapped to *Kismet*'s lip postures. *Kismet* obviously has a limited repertoire as it cannot make many of the lip movements that humans do. For instance, it cannot protrude its lips (important for *sh* and *ch* sounds), nor does it have a tongue (important for *th* sounds), nor teeth. However, computer-animated lip synchronization often maps the 45 distinct English phonemes onto a much more restricted set of visually distinguishable lip postures; 18 is preferred [16]. For cartoon characters, a subset of 10 lip and jaw postures is enough for reasonable artistic conveyance [17]. *Kismet*'s 10 lip postures tend toward the absolute minimal set specified by Fleming and Dobs [17], but is reasonable given its physical appearance. As the robot speaks, new lip posture targets are specified at 33 Hz. Since the phonemes do not change this quickly, many of the phonemes

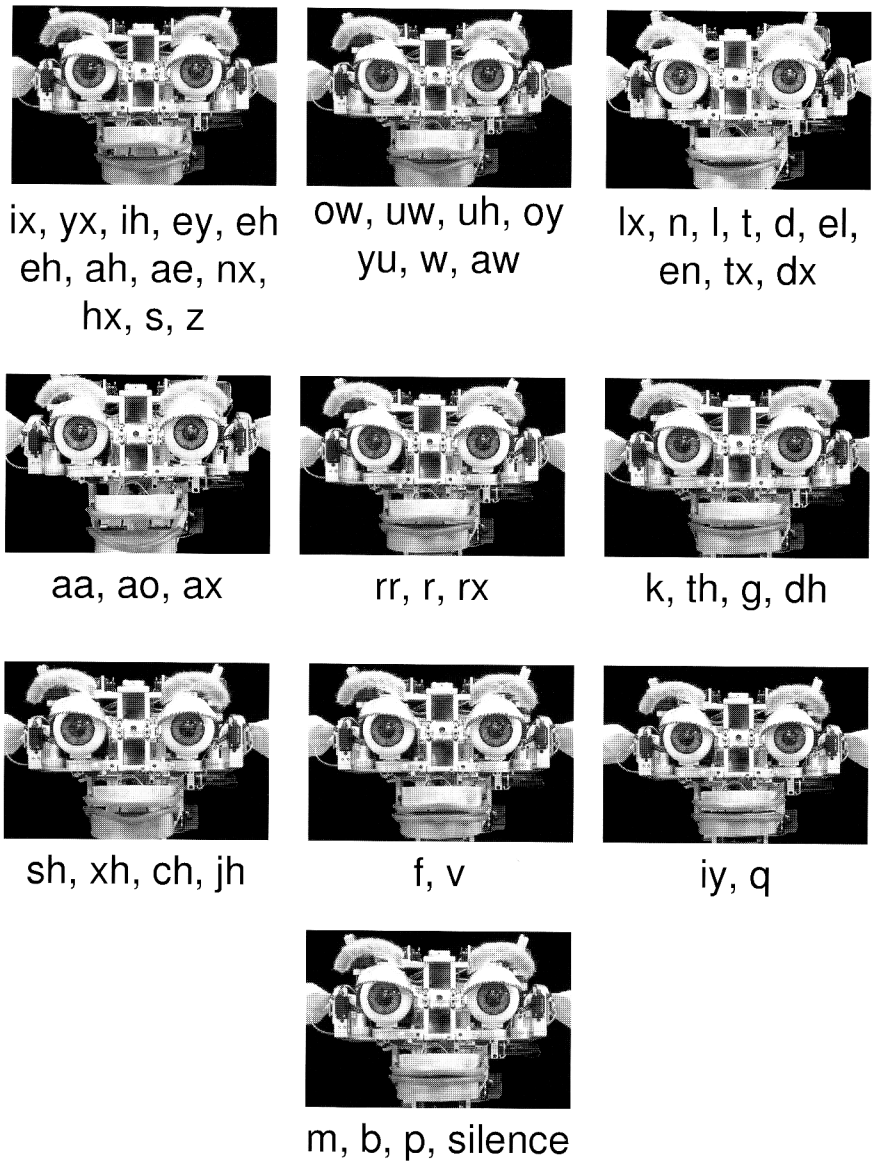

Figure 4. Mapping of phonemes to *Kismet*'s lip postures.

repeat. There is an inherent limit in how fast *Kismet*'s lip and jaw motors can move to the next commanded, so the challenge of co-articulation is somewhat addressed of by the physics of the motors and mechanism.

Lip synchronization is only part of the equation, however. Faces are not completely still when speaking, but move in synchrony to provide emphasis along with the speech. Using the energy of the speech signal to animate *Kismet*'s face (along with the lips and jaw) greatly enhances the impression that *Kismet* 'means' what it says. For *Kismet*, the energy of the speech signal influences the movement of

its eyelids and ears. Larger speech amplitudes result in a proportional widening of the eyes and downward pulse of the ears. This adds a nice degree of facial emphasis to accompany the stress of the vocalization.

Since the speech signal influences facial animation, the emotional correlates of facial posture must be blended with the animation arising from speech. The emotional expression establishes the baseline facial posture about which all facial animation moves. The current emotional state also influences the speed with which the facial actuators move (lower arousal results in slower movements, higher arousal results in quicker movements). In addition, emotions that correspond to higher arousal produce more energetic speech, resulting in bigger amplitude swings about the expression baseline. Similarly, emotions that correspond to lower arousal produce less energetic speech, which results in smaller amplitudes. The end product is a highly expressive and coordinated movement of face with voice. For instance, angry speech is accompanied by large and quick twitchy movements of the ears eyelids. This undeniably conveys agitation and irritation. In contrast, sad speech is accompanied by slow, droopy, listless movements of the ears and eyelids. This conveys a forlorn quality that often evokes sympathy from the human observer.

6. 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 have already found that *Kismet*'s facial expression and body posture help to resolve the ambiguities encountered through voice alone [6]. The ability for *Kismet* to synchronize both lip movement and facial animation in real-time substantially contributes to it delivery of expressive speech.

Acknowledgements

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

REFERENCES

1. R. Picard, *Affective Computation*. MIT Press, Cambridge, MA (1997).
2. C. Breazeal and L. Aryananda, Recognition of affective communicative intent in robot-directed speech, *Auton. Robots* **12** (1), 83–104 (2000).
3. D. Roy and A. Pentland, Automatic spoken affect analysis and classification, in: *Proc. Int. Conf. on Automatic Face and Gesture Recognition (ICAFGR96)*, Killington, VT (1996).

4. L. Chen and T. Huang, Multimodal human emotion/expression recognition, in: *Proc. Int. Conf. on Automatic Face and Gesture Recognition (ICAFGR98)*, Nara, Japan, pp. 366–371 (1998).
5. S. Y. Yoon, B. Blumberg and G. Schneider, Motivation driven learning for interactive synthetic characters, in: *Proceedings of Autonomous Agents (Agents 2000)*, Barcelona, Spain (2000).
6. C. Breazeal, Socialbe Machines: expressive social exchange between humans and robots. PhD thesis, MIT Dept. EECS, Cambridge, MA (2000).
7. I. Murray and L. Arnott, Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, *J. Acoust. Soc. Am.* **93** (2), 1097–1108 (1993).
8. J. Cahn, Generating expression in synthesized speech, MS thesis, MIT Media Lab, Cambridge, MA, 1990.
9. J. Velasquez, Modeling emotions and other motivations in synthetic agents, in: *Proc. AAAI'97*, Providence, RI, pp. 10–15 (1997).
10. D. Canamero, Modeling motivations and emotions as a basis for intelligent behavior, in: *Proc. Agents'97*, Marina Dcl Rey, CA, pp. 148–155 (1997).
11. F. Hara, Personality characterization of animate face robot through interactive communication with human, in: *Proc. IARP98*, Tsukuba, p. IV-1 (1998).
12. P. Blair, *Animation: Learning How to Draw Animated Cartoons*. Walter T. Foster Art Books, Laguna Beach, CA (1949).
13. F. Parke, Computer generated animation of faces, Technical Report UTEC-CSc-72-120, University of Utah, Salt Lake City (1972).
14. R. Madsen, *Animated Film: Concepts, Methods, Uses*. Interland, New York, NY (1969).
15. K. Waters and T. Levergood, DECface: an automatic lip synchronization algorithm for synthetic faces, CRL 94/4, DEC Cambridge Research Laboratory, Cambridge, MA (1993).
16. F. Parke and K. Waters, *Computer Facial Animation*. A. K. Peters, Wellesley, MA (1996).
17. B. Fleming and D. Dobbs, *Animating Facial Features and Expressions*. Charles River Media, Rockland, MA (1999).

ABOUT THE AUTHOR

Cynthia Breazeal directs the Robotic Presence Group at the MIT Media Lab. She has developed numerous autonomous robots, from planetary micro-rovers, to upper-torso humanoid robots, to highly expressive robotic faces. Always inspired by the behavior of living systems, scientific models and theories as well as artistic insights factor heavily into the hardware and software design of her robotic creations. Her current interests focus on social interaction and socially situated learning between people and life-like robots. She received her ScD and SM degrees from MIT in the department of Electrical Engineering and Computer Science with specialization in robotics and artificial intelligence.