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 sociaible robot like Kismet. If done properly, Kismet could then use its emotive vocalizations to convey disagreement, 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 detri ment 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.

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

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Fast</th>
<th>Average</th>
<th>Slower</th>
<th>Slightly Faster</th>
<th>Slightly Slower</th>
<th>Much Faster</th>
<th>Much Slower</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch</td>
<td>Very High</td>
<td>High</td>
<td>Low</td>
<td>Very Low</td>
<td>Low</td>
<td>Very High</td>
<td>Low</td>
</tr>
<tr>
<td>Loudness</td>
<td>Normal</td>
<td>Low</td>
<td>High</td>
<td>Normal</td>
<td>High</td>
<td>Normal</td>
<td>Low</td>
</tr>
<tr>
<td>Voice quality</td>
<td>Smooth</td>
<td>Rough</td>
<td>Smooth</td>
<td>Rough</td>
<td>Smooth</td>
<td>Rough</td>
<td></td>
</tr>
<tr>
<td>Pitch changes</td>
<td>Normal</td>
<td>Low</td>
<td>High</td>
<td>Normal</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
</tbody>
</table>

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...
include accent shape, average tor for affective information. The pitch-related parameters contour of the speech signal, which is the primary contribu-trajectory of the fundamental frequency, speech signal. The pitch-related parameters affect the pitch contour of the spoken utterance. The pitch contour is the pitch, timing, voice quality, and articulation aspects of the The following six pitch parameters influence the pitch

alnal

sortment of vocal affect parameters (VAP) that alter the system, heart rate, and blood pressure. There are an as-sulate the respiratory system, larynx, vocal tract, muscular

settings.

effective qualities, the vocal affect parameters, and the synthesizer

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.

<table>
<thead>
<tr>
<th>DECtalk Synthesizer Setting</th>
<th>DECtalk Symbol</th>
<th>norm</th>
<th>Controlling Vocal Affect Parameter(s)</th>
<th>Percent of Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>average pitch</td>
<td>ap</td>
<td>21</td>
<td>average pitch</td>
<td>1</td>
</tr>
<tr>
<td>assertiveness</td>
<td>as</td>
<td>65</td>
<td>final lowering contour direction</td>
<td>2</td>
</tr>
<tr>
<td>baseline fall</td>
<td>bl</td>
<td>0</td>
<td>contour direction</td>
<td>-5</td>
</tr>
<tr>
<td>breathiness</td>
<td>br</td>
<td>46</td>
<td>breathiness</td>
<td>1</td>
</tr>
<tr>
<td>comma pause</td>
<td>xp</td>
<td>238</td>
<td>speech rate</td>
<td>-1</td>
</tr>
<tr>
<td>gain of fricition</td>
<td>gf</td>
<td>8</td>
<td>precision of articulation</td>
<td>1</td>
</tr>
<tr>
<td>gain of aspiration</td>
<td>gh</td>
<td>0.93</td>
<td>precision of articulation</td>
<td>1</td>
</tr>
<tr>
<td>gain of voicing</td>
<td>gv</td>
<td>0.76</td>
<td>loudness</td>
<td>6</td>
</tr>
<tr>
<td>hot rise</td>
<td>hr</td>
<td>2</td>
<td>reference line</td>
<td>1</td>
</tr>
<tr>
<td>laryngealization</td>
<td>la</td>
<td>0.93</td>
<td>laryngealization</td>
<td>1</td>
</tr>
<tr>
<td>loudness</td>
<td>lo</td>
<td>5</td>
<td>loudness</td>
<td>1</td>
</tr>
<tr>
<td>lax breathiness</td>
<td>lx</td>
<td>75</td>
<td>breathiness</td>
<td>1</td>
</tr>
<tr>
<td>period pause</td>
<td>pp</td>
<td>0.06</td>
<td>speech rate</td>
<td>-1</td>
</tr>
<tr>
<td>pitch range</td>
<td>pr</td>
<td>3</td>
<td>pitch range</td>
<td>1</td>
</tr>
<tr>
<td>quickness</td>
<td>qu</td>
<td>5</td>
<td>pitch discontinuity</td>
<td>-1</td>
</tr>
<tr>
<td>speech rate</td>
<td>sa</td>
<td>2</td>
<td>speech rate</td>
<td>1</td>
</tr>
<tr>
<td>richness</td>
<td>ri</td>
<td>4</td>
<td>brilliance</td>
<td>1</td>
</tr>
<tr>
<td>smoothness</td>
<td>sm</td>
<td>0.65</td>
<td>brilliance</td>
<td>-1</td>
</tr>
<tr>
<td>stress rise</td>
<td>sr</td>
<td>250</td>
<td>accent shape</td>
<td>5</td>
</tr>
</tbody>
</table>

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 **pitch 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:

- **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 friction noise, whereas greater enunciation for consonants results in increased friction. 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.

![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.](image)

### B. Mapping VAPs to Synthesizer Settings

This section presents the equations that map the vocal affect parameters to synthesizer settings. 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:

$$ PP_i = \frac{VAP_{value}}{VAP_{max} - VAP_{min}} $$

$VAP_i$ 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 ($WC_{ji}$) 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 ($SF_{ji}$). 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, $SF_{ji}$ of $VAP_i$:

- If $SF_{ji} \geq 0$
  \[ WC_{ji} = PP_i \times SF_{ji} \]
- If $SF_{ji} \leq 0$
  \[ WC_{ji} = (1 - PP_i) \times (-SF_{ji}) \]

$$ SS_j = \sum_i WC_{ji} $$

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 $SS_{j,final} = SS_j - \text{norm}$
  - If $SS_{j,final} \geq 0$
    \[ SS_{j,final} = SS_{j,final} + (2 \times SS_{j,final} \times (SS_{j,final} - SS_{j,min})) \]
  - If $SS_{j,final} \leq 0$
    \[ SS_{j,final} = SS_{j,final} + (2 \times SS_{j,final} \times (SS_{j,max} - SS_{j,final})) \]

### 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.

<table>
<thead>
<tr>
<th></th>
<th>Anger</th>
<th>Calm</th>
<th>Fear</th>
<th>Happy</th>
<th>Sad</th>
<th>Surprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Range</td>
<td>652.0</td>
<td>564.5</td>
<td>602.0</td>
<td>662.0</td>
<td>510.0</td>
<td>630.0</td>
</tr>
<tr>
<td>Pitch Variance</td>
<td>343.5</td>
<td>269.5</td>
<td>339.5</td>
<td>289.5</td>
<td>280.0</td>
<td>300.0</td>
</tr>
<tr>
<td>Vocal Intensity</td>
<td>22.0</td>
<td>24.0</td>
<td>18.0</td>
<td>20.0</td>
<td>21.0</td>
<td>22.0</td>
</tr>
</tbody>
</table>

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.” The first two test phrases were selected because they follow: *gust, 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.

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.

Three of the 18 utterances were selected to illustrate how 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 slightly 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.

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.”

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.

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


