From Sad to Glad: Emotional Computer Voices

Janet E. Cahn
MIT Media Technology Lab
Cambridge MA 02139

Synthesized English speech is readily distinguished from human speech on the basis of inappropriate intonation and insufficient expressiveness. This is a drawback for conversational computer systems. Intonation is the carrier of emphasis or de-emphasis, serving to clarify meaning for the spoken word much as variations in typeface and punctuation do for the written word. Expressiveness is not tied to word or phrase meaning but is global in scope. It provides the context in which the intonation occurs, and reveals the speaker's intentions and general mental state. In synthesized speech, intonation makes the message easier to understand; enhanced expressiveness contributes to dramatic effect, making the message easier to listen to.

1 Acoustic Correlates of Emotion

Expressiveness and intonation in English are intertwined. Intonation is applied within those acoustic parameters which vary in response to emotional state or attitude. The physiological effects of emotion — arousal, depression — necessarily affects the speech apparatus, and thereby, the speech output. An increase in arousal, as for fear or anger, is characterized by increased heart rate and blood pressure, changes in the rate, depth and pattern of respiratory movements, drying of the mouth and occasional muscle tremor. Speech is fast and loud, with greatest energy in the higher frequencies. Pitch range expands, median pitch is higher than in normal speech, and fluctuations in the pitch contour increase. Enunciation also becomes more precise. Conversely, a decrease in systemic arousal as occurs for relaxation or grief, is characterized by decreased heart rate and blood pressure and an increase in salivation. Speech is slow and low–pitched, high frequencies are weak and enunciation loses precision.

2 Implementation

The most perceptually significant influence of emotions on speech occurs for melody and rhythm. The melody of synthetic speech can be varied for expressiveness by changing its median pitch, expanding or compressing the pitch range, and by shifting the whole pitch range higher or lower. Speech rhythm can be varied by changing the speech rate, pause frequency and location, the regularity of spacing between stressed syllables, and the relative duration of vowels and consonants.

Emotional state also affects speech in the areas of precision of articulation, and voice quality. Precision of articulation can be varied by introducing or removing allophones to effect degrees of clipped or slurred speech. Voice quality can be varied by changes in intensity, by introducing or limiting voicing irregularities such as vocal jitter, and by altering the spectra, particularly the relative strengths of high and low frequency energy in the signal.

3 Preliminary Results

Preliminary attempts to vary the same speech parameters for synthetic speech have produced some utterances that sound distinct from standard synthesis, but with no recognizable affect, and others wherein emotions — for example, fear, irritation, and enthusiasm — are readily discerned. The heart of the initial research is here, in identifying and classifying the perceptual results of speech attributes varied individually and in groups, and in determining the overall sufficiency of the set of aforementioned speech attributes for the production of affect.

4 Synthesizer Limitations

There is a question as to whether today's synthesizers can accommodate the needs of even the lowest level specification of affect. Fortunately, in the key areas of melody and pitch, variation is effected by simply manipulating phoneme duration and frequency. Problems lie in how well synthesizers can handle these phoneme instructions when received at relatively short intervals, say twenty milliseconds. In some instances, phoneme quality suffers. At other times speech output stops while the instructions are processed.

It is not clear whether it is the initial translation of instruction strings to machine instructions that causes the occasional overloads, or the subsequent carrying out of the actual instructions that is at fault. At least, the instruction strings could be compressed by not tying pitch and duration instructions to phonemes, but to larger units - syllables or words. One means of compression is to divide the information into parts – one, a string of phonemes and start times for each syllable, the other, a string containing only frequency local minima and maxima. A pitch contour is derived by interpolation between the maxima and minima, and phonemes are matched to frequencies based on when each occurs.

As long as all values are located in time, this paradigm can be adopted for other speech parameter settings that are currently static for a whole utterance. For example, local maxima and minima for intensity values could be sent to the synthesizer along with the phoneme and pitch specifications, so that loudness would change over the course of one utterance. Voice quality changes could be similarly specified. Expansion of the kinds of speech parameters that can be dynamically controlled, coupled with minimal specification requirements, will go a long way towards providing the capacity for synthesizing convincingly expressive speech.