Generation of Affect in Synthesized Speech

Janet E. Cahn
MIT Media Technology Laboratory
20 Ames Street
Cambridge, MA 02139
phone: 617-253-0666

Generation of Affect in Synthesized Speech

1 Introduction

When compared to human speech, synthesized speech is distinguished by insufficient *intelligibility*, inappropriate *prosody* and inadequate *expressiveness*. These are serious drawbacks for conversational computer systems. *Intelligibility* is basic — intelligible phonemes are necessary for word recognition. *Prosody* — intonation (melody) and rhythm — clarifies syntax and semantics and aids in discourse flow control. *Expressiveness*, or *affect*, provides information about the speaker's mental state and intent beyond that revealed by word content.

My work explores improvements to the affective component of synthesized speech. It is embodied in the Affect Editor program, which is intended to show that variations in affect can be generated in synthetic speech and to point the way towards improving the recognizability and naturalness of the affect [2]. Its success in generating recognizable affect was confirmed by an experiment in which the intended affect was perceived for the majority of presentations.

Affect should be a concern in speech synthesis for theoretical and practical reasons. Its role in human speech is to provide the context in which utterances should be interpreted and to signal speaker intentions. It should play the same role in synthesized speech. More practically, the addition of controllable affect will allow synthesized speech to be used in any application in which expressiveness is appropriate — for example, tools for the presentation of dramatic material, information–giving systems and synthesizers used by the speech–handicapped.

2 Prerequisites

To incorporate affect into synthesized speech, one must identify the acoustic correlates of emotions and choose a representation that allows for systematic control of these correlates. Both are discussed in this section.

2.1 Speech correlates of emotion

Speech correlates of emotion have been identified by studies of the speech signal [4, 5, 9] and of perceptual responses to emotional speech [3, 8]. The primary conveyers of affect are fundamental frequency (F0) and duration. Much of the associated speech phenomena are explained by the effect of emotion on physiology. With the arousal of the sympathetic nervous system — as for fear, anger or joy — heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. Speech is correspondingly loud, fast and enunciated and has much high frequency energy. With the arousal of the parasympathetic nervous system — as for boredom or sadness — heart rate and blood pressure decrease and salivation increases. Speech is slow and low–pitched and high frequency energy is weak [10].

2.2 Representation of the effect of emotion on speech

The representation of the speech correlates of emotion can proceed from a speaker model or an acoustic model. The first is the more generative approach — the effects of emotion on physiology and hence, on speech, are derived from the representation of the speaker's mental state and intentions. The second describes primarily what the listener hears. Of the two, the acoustic model is the simpler and requires less knowledge overall. Moreover, since perceptual parameters are explicit in the model, they can be easily manipulated to test perceptual responses. This is the model that is incorporated into the Affect Editor.

The parameters of the acoustic model are grouped into four categories — pitch, timing, voice quality and articulation. The pitch parameters describe features of F0. The timing parameters control speed and rhythm (the combination of word stress and silence). Often pitch and timing parameters describe features of words or phrases. In contrast, the voice quality parameters affect features of the speech signal as a whole. Articulation parameters control variations in enunciation, from slurred to precise.

The pitch parameters are:

- *Accent shape*: the steepness of the intonation contour at the site of a *pitch accented*[1] syllable.
- *Average pitch*: the average F0 for the utterance.
- *Contour slope*: the slope (increasing, level or decreasing) of the F0 contour.
- *Final lowering*: the steepness of the F0 decrease at the end of falling contours, or of the rise at the end of rising contours (as for yes-no questions).
- *Pitch range*: the distance between the speaker's lowest and highest F0 for the utterance.
- *Reference line*: a term borrowed from work on generative intonation [1]. It specifies the F0 that is returned to after a high or low pitch excursion.

The timing parameters are:

[1] A word stressed with noticeably high or low F0 is a *pitch accented* word. This significant pitch occurs within the syllable carrying the greatest lexical stress [7].

- *Exaggeration*: the degree to which pitch accented words receive exaggerated duration as a means of further emphasis. (The *exaggeration* is currently unimplemented because of synthesizer introduced side effects.)
- *Fluent pauses*: the frequency of pausing between syntactic and semantic units.
- *Hesitation pauses*: the frequency of pausing within a syntactic or semantic unit
- *Speech rate*: the rate of speech. The *speech rate* affects the number of words spoken per minute and the pause lengths.
- *Stress frequency*: the percent of stressed words to stressable words for an utterance.

The voice quality parameters are:

- *Breathiness*: the amount of frication noise that may be co–present with non–fricative phonemes (vowels, for example).
- *Brilliance*: the ratio of low to high frequency energy.
- *Laryngealization*: vocal fold vibration unaccompanied by air flow across the vocal cords. The speech of older speakers is often laryngealized.
- *Loudness*: loudness.
- *Pause discontinuity*: the smoothness or abruptness of a pause onset.
- *Pitch discontinuity*: the distance between successive F0 values in the intonation contour.
- *Tremor*: regularities between successive glottal pulses. It is not a Dectalk3 parameter, so currently has no effect on the output.

The sole articulation parameter is:

- *Precision*: the degree of slurring or enunciation. Perhaps *precision* should be expanded so that it is specified separately for phonemes distinguished by means of production or place of articulation.

The acoustical model has two significant features. First, the parameter values are quantified on an abstract scale that is centered around 0 and ranges from -10 to 10. This allows simple comparisons between the effects of different emotions. New affective colorations can be created by changing the parameter values. Additionally, perceptual thresholds may be tested via incremental changes.

Note that parameter values are arranged such that the mid–range value (0) for the range quantifies a neutral affect, while the extreme values of 10 and -10 represent a parameter's maximum and minimum influence, respectively. This makes the implementation of affect quantifiers straightforward. Moving the values away from their neutral settings corresponds to more of an affective coloration, while moving the values closer corresponds to less.

2.2.1 Input

The input to the Affect Editor is an annotated utterance, meant to represent the plausible output of a text generation program. The phrases in the utterance are grouped according to their roles, as per case frame descriptions. For example:

[[SUBJ I thought] [OBJ you [ACTION really meant it]]]

Additionally, each word is annotated with its part of speech and the probability of its receiving significant stress. The Affect Editor assigns word stress and enunciation from word annotation and determines pause locations and the shape of the terminal contour from phrase annotation.

2.3 Discussion

The Affect Editor implements a transfer function from an acoustical description of emotional speech to synthesized speech. It is a tool with which the effects of different emotions on speech can be quantified and correlations between acoustic parameters can be investigated.

3 Experiment

A simple experiment was performed in which five affectively neutral sentences were synthesized with six different affects and presented in random order to twenty–eight subjects. The subjects heard these sentences —

```
I'm almost finished.
I saw your name in the paper.
I thought you really meant it.
I'm going to the city.
Look at that picture.
```

— presented for angry, fearful, disgusted, surprised, sad and glad affective coloring.

The hypothesis had two parts: 1) that the affect conveyed by the voice would be recognized despite the sentence semantics and 2) that semantically and acoustically similar emotions would be confused. In fact, sentence semantics *did* color the judgments. For example, the sentence *"I thought you really meant it."* was rarely perceived as glad, though often, instead, as surprised. Of all the emotions, sadness was the most consistently recognized. This makes sense because its characteristics — soft, slow, halting speech, with minimal high frequency energy — are the most distinct. The other emotions were recognized a little over 50% of the presentations, and were mistaken for similar emotions (anger and disgust, gladness and surprise) for an additional 20%.

The number of correct recognitions was significantly above chance (at 17% correct). Errors were not random, but followed the pattern of errors made when identifying affect in human speech. The experiment answered broadly, in the affirmative, the question of whether recognizable affect could be generated in synthesized speech.

3.1 Synthesizer considerations and effects

Because the acoustic representation is synthesizer independent its parameters must be interpreted for each synthesizer it drives. The mapping for the Dectalk3 – the only synthesizer used so far — involves both one–to–many and many–to–one mappings from the acoustic parameters to the synthesizer settings. The parameters not represented in the Dectalk's parameter set are instead implemented in software. Thus, a contour slope up or down is approximated by assigning a high F0 to the word at either end of the utterance, pauses are added by inserting a silence character, the smoothness or abruptness of the pause onset is effected by inserting phonemes prior to the silence and precision of articulation is achieved by phoneme substitutions or additions.

The Dectalk3 was chosen because it allows control over aspects of pitch, timing, voice quality and articulation. However, some of its limitations make it unclear whether an emotion is poorly specified or correctly specified but poorly reproduced. The limitations are of two kinds — side effects and limited capabilities. A side effect of expressing a word in phonemic form is that it often has a lower F0 than when it is specified with text. Word stress markings should cause primarily local perturbations but instead sometimes affect the intonation contour before and after the word such that subsequent word stresses are unrealized. Finally, as the average pitch is raised the voice begins to sound like a different speaker, instead of a speaker raising her voice. This side effect occurs because the pitch range and average pitch values are interdependent but should instead be independent.

Many of these side effects can be remedied by implementing an intonational description system with primarily local effects (such as the two tone annotation developed by Pierrehumbert and colleagues [7, 6, 1]). Affect synthesis would be made easier by the addition of the features implemented in software in the Affect Editor, particularly the ability to specify the precision of articulation and the overall pitch contour slope.

4 Conclusions

This work shows that recognizable and even natural–sounding affect can be produced by imitating in synthesized speech the effects of emotion on human speech. More importantly, it is a tool for exploring what is needed in an affect generating system. Its effectiveness would be increased with synthesizers that more accurately reproduce the Affect Editor specifications.

References

[1] Mark Anderson, Janet Pierrehumbert, and Mark Liberman. Synthesis by Rule of English Intonation Patterns. In *Proceedings of the Conference on Acoustics, Speech, and Signal Processing*, page 2.8.1 to 2.8.4, 1984.

[2] Janet E. Cahn. Generating Expression in Synthesized Speech. Master's thesis, Massachusetts Institute of Technology, May 1989. Unpublished.

[3] Joel Davitz. *The Communication of Emotional Meaning.* McGraw-Hill, 1964.

[4] G. Fairbanks. Recent experimental investigations of vocal pitch in speech. *Journal of the Acoustic Society of America*, (11):457–466, 1940.

[5] G. Fairbanks and W. Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotions. *Speech Monographs*, 6:87–104, 1939.

[6] Mark Liberman and Alan Prince. On Stress and Linguistic Rhythm. *Linguistic Inquiry*, 8(2):249–336, 1977.

[7] Janet B. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology, 1980.

[8] Klaus Scherer. Speech and emotional states. In John K. Darby, editor, *Speech Evaluation in Psychiatry*, pages 189–220. Grune and Stratton, Inc., 1981.

[9] Carl E. Williams and Kenneth N. Stevens. On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Medicine*, 40(12):1369–1372, December 1969.

[10] Carl E. Williams and Kenneth N. Stevens. Emotions and Speech: Some Acoustical Correlates. *Journal of the Acoustic Society of America*, 52(4 (Part 2)):1238–1250, 1972.