TECHNIQUES FOR SYNTHESIZING VISIBLE, EMOTIONAL SPEECH

Caroline Henton

Speech And Voice Information Analysis *and*
Linguistic Research Center, University of California Santa Cruz

ABSTRACT — Attempts are described to synthesize visible speech in real-time on a Macintosh™ personal computer, and to enable the user to color the text of the speech to be synthesized emotionally, according to the user's wishes, the representation of the text, or the semantics of the utterance. Animated visible speech will be demonstrated, using a variety of on-screen agents.

INTRODUCTION

Diverse research in ethology, psychology, speech therapy, interpersonal communications and human-machine interface design indicates that facial expressions enhance communication. Facial expressions convey emotion (Ekman and Friesen, 1986) and other communicative signals (Chovil, 1991). A multimodal form of interaction with a computer on-screen agent is likely to improve the quality of the dialogue, both in terms of intelligibility and with regard to user-friendliness (Pelachaud et al., 1993).

This paper describes attempts to enhance the user-dialogue with an on-screen agent, specifically by providing synthetic visible speech with more accurate, pleasing articulation, in real-time on a Macintosh™ personal computer (1), and by enabling emotional 'coloring' of text to be synthesized. Details of the creation of units of visible speech and their co-ordination with animated faces are given in Henton and Litwinowicz (1994). A brief overview of that work is given below. Two further enhancements are outlined for users of text-to-speech systems. The first is the ability to generate vocal emotion in concatenated synthetic speech. The second allows for vocal emotions to be included during the authoring of text for output by the text-to-speech system. The animated visible emotional speech will be demonstrated, in real time, using a variety of on-screen agents and faces.

VISIBLE SPEECH

To make facial animation of on-screen agents realistic, sound output from the synthesizer and facial movements must be synchronized. Previous visual speech systems have used 'visemes', i.e. minimal contrastive units of visible articulation of speech sounds; the number of visemes has typically ranged from 9 to 32 for American English. Animations based on transitions from one viseme to another viseme have been choppy, inaccurate and insufficiently plastic. In contrast, concatenative speech synthesis systems, such as that embodied in the Apple Macintosh MacinTalkPro2™ (1) text-to-speech system, require a large inventory of diphone units (about 1800) from which to create utterances. Two improvements are therefore needed: an expansion beyond simple visemes, and a reduction of the number of diphones so that they may be mapped to the facial images in real-time on a personal computer. If this can be effected, the speech transitional movements for a speaker image might be reflected more accurately. To this end, visible and articulatory archiphones and diphone 'aliases' were formalized, using standard phonological distinctive features, and a system of *disemes* was created. Disemes may be defined as beginning during one viseme (phone) in an archiphonic family and ending somewhere during the following viseme (phone), in another archiphonic family (Henton, 1994). The transitioning that occurs in a diseme is generally not a linear function, but rather depicts the varying rates of articulatory imaging which occur in a human speaker. In this way, many of the transitions that occur in diphones for General American English can be visually depicted by the same diseme, due to their similarity in lip, teeth, and tongue image positions. For further details about the animation techniques, see Henton and Litwinowicz (1994).

EMOTIONS IN SYNTHETIC SPEECH

Synthesized speech has been generally 'neutral' in tone, disinterestedly dull, and lacking in vocal emotionality. In the most parsimonious cases, it has been a monotone. This deficiency is partly accounted for by the default intonation tunes in speech synthesizers. Means have existed to make the synthetic speech sound, for example, happy or angry, but research has been directed primarily towards maximizing intelligibility rather than naturalness, or variety. Research into techniques for adding emotional affect to synthetic speech has concentrated solely on parametric synthesizers (e.g. Cahn, 1990; Murray, Arnott & Newell, 1991). Cahn's study (1990) produced mixed results and remains inconclusive about "the perception of affect in speech" (p.139). There is widespread acknowledgment that perception of synthesized emotions is not predictable and may not yet be a precise science (cf. Scherer, 1984).

If computer memory and processing speed were unlimited, a possible method for creating vocal emotions might be to simply store words spoken by a human being in varying emotional ways. For applications to run on many personal computers, this is impractical. Rather than being stored, emotions have to be synthesized on-line and in real-time. In parametric synthesizers, there may be as many as thirty basic acoustic controls available for altering pitch, duration and voice quality. These include, e.g., separate control of formants' values and bandwidths; pitch movements on, and duration of, individual segments; breathiness; smoothness; richness; assertiveness; etc. Precision of articulation of individual segments (e.g., fully released stops, degree of vowel reduction), which is controllable in DECtalk, can also contribute to the perception of emotions, such as tenderness and irony. These parameters may be manipulated to create voice personalities; DECtalk is supplied with nine different 'Voices' or personalities. It should be noted that intensity (volume) is not controllable within an utterance in DECtalk.

Limitations of a concatenative synthesizer

In diphone-concatenative speech synthesizers, such as MacinTalkPro 2, control of individual acoustic features is severely limited. Firstly, it is not possible to alter the voice quality of the speaker, since the speech is created from the recording of a live speaker (who has their individual voice quality) speaking in one (neutral) vocal mode, and parameters for manipulating positions of the vocal folds are not included in this type of synthesizer. Secondly, precision of articulation of individual segments is not controllable in this concatenative synthesizer. It is nonetheless possible in MacinTalkPro 2 to control the following parameters: Average speaking pitch; Pitch range; Speech rate; Volume; Silence; Pitch movements; Duration. Although 7 parameters are listed, it is nevertheless possible to produce a wide range of emotional affect using the interplay of only 5 parameters – since Speech rate and Duration, and Pitch range and Pitch movements are, respectively, effected by the same acoustic controls. Details for using the parameters with MacinTalkPro 2 are published in the Macintosh Developer Note #5 (1993).

Example emotions and specifications

Table I, below, gives examples of some emotions which have been defined, together with their associated vocal emotion values. It should be remembered that these values were designed to apply to General American English, and users would need different vocal emotion values to be specified for application to other dialects and languages. Nevertheless, the particular values shown are easily modifiable, to allow for differences in cultural interpretations and user/listener perceptions.

The values (and underlying comments) in Table I are relative to the default neutral speech setting for a high-quality female voice in MacinTalkPro 2. For a male voice, the values need to be altered. For example, the default specification for a high-quality male voice in MacinTalkPro 2 might use a pitch mean of 43 and a pitch range of 8 (thus specifying a lower, but more dynamic, range than the female voice of 56; 6). However, in general, neither volume nor speaking rate is sex-specific, and, as such, these values would not need to be altered dramatically when changing the sex of the speaking voice. As for determining values for other vocal emotions when changing to a male speaking voice, these values could merely change as the female voice specifications do, relative to the default specification. It should be noted that in MacinTalkPro 2 the default speech rate is 175 words per minute (wpm) whereas a realistic human speaking rate range is 50-500 wpm.

Using the command set and calculations given in Macintosh Developer Note #5 (1993), the values in Table I are input to the speech synthesizer, The parameters pitch mean and pitch range are represented acoustically in a logarithmic scale in the speech synthesizer. Logarithmic values are converted to a linear scale of integers in the range 0 - 100 for the user's convenience. On this scale, a change of +12 units corresponds to a doubling in frequency, while a change of -12 units corresponds to a halving in frequency. Because pitch mean and pitch range are each represented on a logarithmic scale, the interaction between them is quite sensitive. On this basis, a pmod value of 6 will produce a markedly different perceptual result with a pbas value of 26 than with 56. The range for volume, conversely, is linear; therefore doubling a volume value results in a doubling of the output volume from the synthesizer. Prosodic commands for Baseline Pitch (pbas), Pitch Modulation (pmod), Speaking Rate (rate), Volume (volm), and Silence (slnc), may be applied at all levels of text, i.e., passage, sentence, phrase, word, phoneme, and allophone.

Table I Examples of some vocal emotions defined according to a restricted set of prosodic values. N.b. these values were designed to apply to a female voice speaking General American English, only.

Emotion	Pitch Mean / Range (pbas)/(pmod)	Volume (volm)	Speaking Rate (rate)
Default (normal)	56; 6 (neutral and narrow)	0.5 (neutral)	175 neutral
Angry1 (threat)	35; 18 (low and narrow)	0.3 (low)	125 (slow)
Angry2 (frustration)	80; 28 (high and wide)	0.7 (high)	230 (fast)
Happy	65; 30 (neutral and wide)	0.6 (neutral)	185 (medium)
Curious	48; 18 (neutral and narrow)	0.8 (high)	220 (fast)
Sad	40; 18 (low and narrow)	0.2 (low)	130 (slow)
Emphasis	55; 2 (neutral and narrow)	0.8 (high)	120 (slow)
Bored	45; 8 (neutral and narrow)	0.35 (low)	195 (medium)

The following examples show the results of applying different vocal emotions to different portions of text. The first scenario shows the result of merely inputting the text into the text-to-speech system and using the default vocal emotion parameters for female voices. In this scene, the portions of text in italics indicate speech by the car repair-shop employee while the rest of the text indicates the car owner. The portions in double brackets indicate the speech synthesizer parameters; and the portions of text in single brackets are merely comments added for clarification here.

1. [Default] [[pbas 56; pmod 6; rate 175; volm 0.5]] Is my car ready? *Sorry, we're closing for the weekend.* What? I was promised it would be done today. I want to know what you're going to do to provide me with transportation for the weekend!

With only the default prosodic values in place, MacinTalkPro 2 could play this scenario through a loudspeaker, however, it would be hard to distinguish the two speakers in the conversation, and the interchange might sound somewhat robotic owing to the lack of vocal emotion. After the application of vocal emotion parameters (either through use of the graphical user interface, direct textual insertion, or other automatic means of applying the defined vocal emotion parameters), the text might look like this:

2. [Default] [[pbas 56; pmod 6; rate 175; volm 0.5]] Is my car ready? [Disinterested] [[pbas 55; pmod 5; rate 170; volm 0.5]] *Sorry, we're closing for the weekend.* [Angry1] [[pbas 35; pmod 18; rate 125; volm 0.3]] What? I was promised it would be done today. [Angry2] [[pbas 80; pmod 28; rate 230; volm 0.7]] I want to know what you're going to do to provide me with transportation for the weekend!

This second scenario thus provides the speech synthesizer with speech parameters that will result in speech output having vocal emotion. Two varieties of 'Anger' are suggested; this emotion has been shown to have two distinct manifestations in speech (Frick, 1986). The first ('Angry1') may be heard as 'cold' anger, a form of controlled threat; the second ('Angry2') is 'hot' anger, being louder, faster, more dynamic and uncontrolled.

Individual words within a passage can receive only one type of modification, specifically where additional emphasis [[emph]] on a single (following) word is achieved by a rise in the pitch and a lengthening of the vowels. Both [[emph]] and [[slnc]] apply only to the following word string, and do not require resetting, or toggling off:

```
                [[pbas 56; pmod 6; rate 175; volm 0.5]]
        This is a [[emph +]] beautiful [[slnc 30]] morning. [[rate 140; volm 0.4]] The sun is piercing
        the sky between [[rate 150; volm 0.6]] black [[rset]] clouds that cling to the mountains' crest.
```

MacinTalkPro 2 also gives the user access to phonemes, the minimal contrastive units of speech. The exact specification of the phonemes used for General American English is not needed here. Modifications to individual phonemes within a passage can be achieved by first entering the phonemic Input Mode [[inpt PHON]] and then adding prosodic inflection controls to the basic phoneme symbols, as illustrated below for the word "anticipation" in the phrase "Anticipation is all.":

```
        [[inpt PHON]] /2AEn=t2IH=sIX=p1>/EY=S/IXn [[inpt TEXT]] is all.
```

The pronunciation of the word "anticipation" could be perceived as being more excited than normal, because of the rising pitch (/) on the first, penultimate and last syllables, and the increased length (>) of the penultimate syllable. In this notation, syllables are divided by the equals sign (=). It is possible to experiment with synergistic combinations of settings to achieve a given emotional connotation. Inflection Control symbols (/,\,<,>) may be concatenated to provide more exaggerated, cumulative effects. The specific nature of the effect depends on the speech synthesizer, and on its perception by the listener.

VISUAL REPRESENTATION AND MANIPULATION OF SPEECH PARAMETERS

As discussed above, terms used in speech synthesis and the existing prosodic controls are not well suited for authoring emotional prose at a high level. The problem lies not only in the terminology, for example "baseline-pitch", but also in the difficulty of quantifying these terms. To reiterate: choosing numerical values for each of several speech parameters to incorporate vocal emotion into each word spoken would be very tiresome. A more intuitive and faster approach is needed.

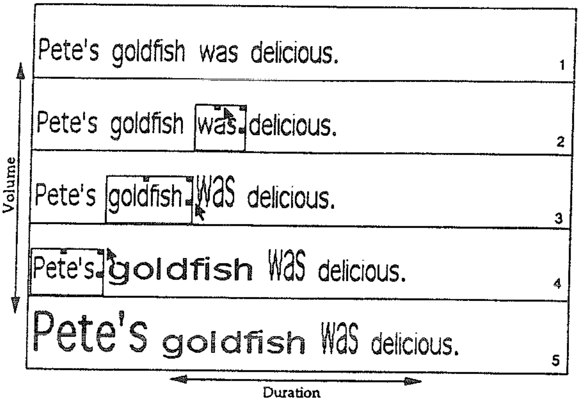
Other graphical interfaces for modification of sound currently exist. For example, commercial products such as SoundEdit®, by Farallon Computing, Inc., provide for manipulation of raw sound waveforms. Manipulation of raw waveforms does not provide a clear intuitive means to specify vocal emotion in synthetic speech because of the lack of clear connection between the displayed waveform and the desired vocal emotion. Simply put, by looking at a waveform of human speech, an acoustically naive user cannot easily ascertain how it (or modifications to it) will sound when played through a loudspeaker, particularly if the user is attempting to provide some sort of vocal emotion to the speech. The graphical interface described below allows the user to visually represent and to control the following vocal characteristics through direct manipulation: volume, duration, pitch variation. By combinations of these three parameters, the user can represent and control some limited vocal emotions.

Volume and duration

The control of volume and duration takes advantage of the two natural spatial axes of a computer display; volume is the vertical axis, duration the horizontal axis. By single clicking on a word in the text to be spoken, that word is selected and available for manipulation. Three sizing grips are presented: one for volume only, one for duration only, and one which allows both volume and duration to be manipulated simultaneously. The word is simply stretched along the axes. The taller a word becomes, the greater volume it will have; likewise, the wider a word becomes, the greater its duration. In Figure 1 the original text is shown, followed by a series of manipulations required to create the resulting text.

Emotion

Colors are used to attach an emotion to a word. Colors may be chosen as they seem appropriate, and to allow for differing cultural implications. For example, in some cultures, yellow may be perceived as happy, while in others yellow may be perceived as angry. For illustration here red will represent angry, and yellow will represent happy. Accordingly, imagine Pete's cat speaking the sentence "Pete's goldfish was delicious". The user would highlight this sentence in standard manner, and select 'Happy' from a range of emotions. The selected text would then turn yellow, the change in color being, of course, independent from its other attributes (its volume and duration). Pete's 'Angry' reply would be shown in red. An emotion called 'Normal' is associated with black and is the default. A screen shot of the editor appears in Figure 2.



Pete's goldfish was delicious. 1

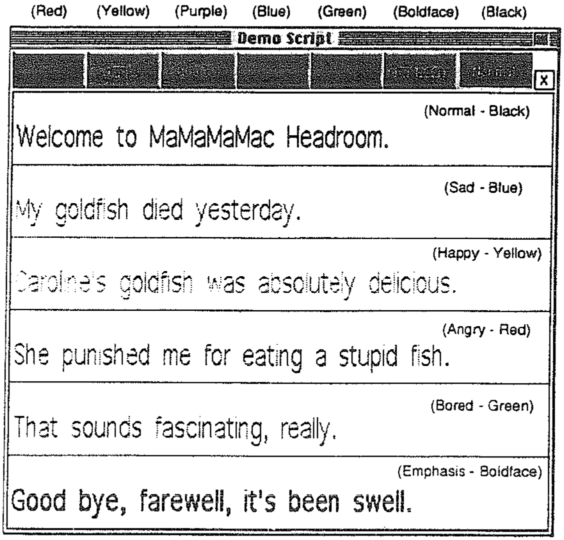
Pete's goldfish was delicious. 2

Pete's goldfish was delicious. 3

Pete's goldfish was delicious. 4

Pete's goldfish was delicious. 5

Figure 1. Original text (line 1) can have a series of manipulations applied to it to create the resulting text (line 5). Three sizing grips are provided (indicated here by the on-screen arrow): one for volume only (line 2), one for duration only (line 3), and one which allows both volume and duration to be manipulated simultaneously (line 4). The selected word is stretched along the axes, labeled here 'Duration' on the x-axis and 'Volume' on the y-axis. Line numbers are included for clarity; they do not appear on the screen.



(Normal - Black)
Welcome to MaMaMaMac Headroom.

(Sad - Blue)
My goldfish died yesterday.

(Happy - Yellow)
Caroline's goldfish was absolutely delicious.

(Angry - Red)
She punished me for eating a stupid fish.

(Bored - Green)
That sounds fascinating, really.

(Emphasis - Boldface)
Good bye, farewell, it's been swell.

Figure 2. Screen shot of the text editor, showing colored buttons used for selecting emotions.

CONCLUSION, LIMITATIONS, AND FURTHER WORK

The techniques for improving the user experience with synthetic speech described here are both relatively simple and novel. However, they are far from ideal or complete. One of the inherent limitations of using disemes in visible speech stems directly from diphones. Phonetic research has shown that anticipatory co-articulation for the lips extends over several segments, and that the acoustic consequences go far beyond one or two phonetic segments ahead. Positions of the lips, teeth, and tongue may thus be altered visibly several hundred milliseconds both before and after the segment currently being spoken by the speech synthesizer. Only when hybrid synthesis produced by a parametric-concatenative system is available can this type of anticipatory modeling be achieved. Similarly, the synthesis system described here is unable to hypo-or hyper-articulate, except by using prosodic means; and this may sometimes result in unnatural trajectories. Nevertheless, initial attempts have been made to overcome difficulties presented by bimodal text-to-speech synthesis. The 2-D on-screen agents are a significant, real-time, computationally low-cost enhancement in human-computer communication.

The model of using limited prosodic parameters to create emotions is needed for non-parametric synthesizers, and the visual representation and manipulation of speech using a graphical editor is intuitive and attractive. Regarding the method of speech synthesis used, any concatenative speech synthesis system will have a set of prosodic controls available for individual manipulation to simulate vocal emotions or personalities; which controls are used, and how, is not previously reported. By combining the acoustic-phonetic parameters and the graphical user interface, a method has been created for authoring emotional prose for use in a text-to-speech system that is both friendly and transparent.

Further work is required to produce more sophisticated speech synthesis, with a greater variety of emotions, and to quantify the gain both in speech intelligibility and in the overall user experience. Using the full dynamics of tracked disemes should produce better diseme animations. Ideally, the goal is to synchronize a hybrid speech synthesizer with even more convincing models of faces; and to provide those faces with fully discriminable vocal emotions .

ENDNOTE 1. Macintosh ® and MacinTalkPro 2® are registered trademarks of Apple Computer, Inc.

REFERENCES

Cahn, Janet E. (1990) *Generating Expression in Synthesized Speech.* Technical Report, M.I.T. Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Digital Equipment Corporation (1985) *DECtalk DTC03 Text-to-Speech System Owner's Manual.* (Maynard, Massachusetts).

Chovil, N. (1991) *Social determinants of facial displays.* Journal of Nonverbal Behavior, 15, 141-154.

Ekman, P. & Friesen, W.V. (1986) *A new pan-cultural facial expression of emotion.* Motivation and Emotion, 10, 159-168.

Frick, R.W. (1986) *The prosodic expression of anger: differentiating threat and frustration.* Aggressive Behavior, 12, 121-128.

Henton, C. (1994) *Beyond visemes: using disemes in synthetic speech with facial animation.* Journal of the Acoustical Society of America, 95, 3010.

Henton, C. and Litwinowicz, P. (1994) *Saying it with feeling: techniques for synthesizing visible, emotional speech.* Proceedings, 2nd. ESCA/IEEE Workshop on Speech Synthesis, 73-76.

Macintosh Developer Note #5 (for Quadra 840 and Centris 660 AV computers). Apple Computer, Inc., Cupertino, CA.

Murray, I.R. and Arnott, J.L. (1993) *Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion.* Journal of the Acoustical Society of America, 93, 1097-1108.

Pelachaud, C., Viaud, M. L. and Yahia, H. (1993) *Rule-structured facial animation system.* IJCAI 1993.

Scherer, K.R. (1984) *Emotion as a multicomponent process: a model and some cross-cultural data.* Review of Personality and Social Psychology, 5, 37-63.