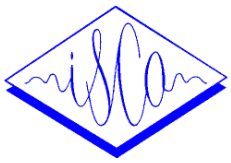
ISCA Archive
http://www.isca-speech.org/archive



First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

ENHANCEMENT OF HUMAN-COMPUTER INTERACTION THROUGH THE SYNTHESIS OF NONVERBAL EXPRESSIONS

Kris Maeda Yasuki Yamashita Yoichi Takebayashi

Toshiba Corporation, Research & Development Center
Saiwai-ku, Kawasaki, 210 Japan

ABSTRACT

This paper addresses the enhancement of human-computer interaction through the synthesis of nonverbal expressions. As a means of improving the naturalness of synthesized speech for the realization of human-computer discourse, study has been focused on the effects of emotion and intention on prosodic features of human speech. Experiments were carried out to observe the time-variant behavior of the fundamental frequency, amplitude, and formant frequencies of nonverbal expressions relative to the intentions and emotional states of the speakers. From the observed data, fundamental relationships were determined and stated in the form of constraints on parameters for a formant synthesizer. Fuzzy inference was used to process these constraints. The method supports the linguistic declaration of constraints in the form of conditionals, and it enables fine-tuning through the adjustment of fuzzy set membership functions.

I. INTRODUCTION

Research efforts in the area of speech synthesis have resulted in the realization of several text-to-speech conversion systems [1]. The synthesized speech is recognizable and is currently used in a number of practical applications; however, most systems do not have the capacity for providing naturalness. The reason is a lack of prosodic features, characteristic of human discourse, which are the consequences of numerous speech production processes including semantics, pragmatics, acoustics, lexical rules, syntax, and phonetics. These processes are interrelated and pose obstacles to endeavors for achieving accurate synthesis of speech. However, researchers in various communities have been gradually uncovering valuable information concerning these processes. Speaker intention and emotional state have been shown to be influential on prosody and discourse, contributing significantly to both utterance and discourse interpretation [2,3,4].

For the purpose of developing human-computer interactive systems, the authors have become increasingly aware of the importance of naturalness to discourse. Interactive systems have requirements which differ from those of conventional text-to-speech systems. Conventional systems support media conversion from the written to the spoken word. Interactive systems must be designed with the objective of eliminating the communication barriers which exist due to the inanimate nature of computers. Thus, synthesized speech should embody warmth and intimacy, being modeled according to natural human discourse.

A fundamental element of human discourse which could greatly enhance synthesized speech for human-computer interaction is the expression of nonverbals. In daily conversation, nonverbals are a common medium for conveying both emotions and intentions. Despite the fact that they are inherent to communication among humans, these utterances have received little attention from the speech research community.

Recognizing the value of nonverbal expressions to natural discourse, the authors are working towards a constraint-based synthesis system which can support and maintain the knowledge necessary for generating nonverbals. Current rule-based techniques are limited to a narrow portion of the speech output domain; the solution lies in utilizing methods of Artificial Intelligence (AI) to maintain and support knowledge representations of human cognition and speech processes.

In this paper, the effects of emotion and intention on speech, including nonverbals, is discussed. Preliminary analysis and synthesis of nonverbals is presented.

II. EMOTION AND INTENTION

Understanding the various working functions of the human mind is a key issue in the development of computers that possess intelligent interactive capabilities. Cognitive scientists and AI researchers are striving toward the modeling and simulation of such functions. One example is the formulation of speaker intentions and their influence on emotional state [5]. Emotions may be represented as positive or negative states of arousal derived from intentional information [6]. If a person fulfills a desired intention, the resulting emotion may be happy; however if that desired intention is not fulfilled, the emotion may be anger or sadness.

Treated independently, intentions strongly influence a person's behavior. A number of processes, including speech production, are controlled as a person strives to fulfill an intention. Actions might be oriented toward achieving a desirable outcome or avoiding an undesirable outcome. Keeping the intention in mind, a person's behavior is adjusted accordingly. Manner of speaking during discourse is just one of the many aspects that are subsequently regulated.

Emotions are regarded as mental states that have direct physiological effects. These effects include facial expressions, which often accompany emotions and can alter the physical structure controlling articulation and acoustics of speech. The position of the lips and tongue contribute to the shape and size of the vocal tract. Thus, speech correlates of emotion may be observed through the median pitch, pitch range, pitch contour shape, pitch variability, formant frequencies, and amplitude.

III. NONVERBAL EXPRESSIONS

Nonverbal expressions, defined as having little lexical content, play a vital role in human communication. They are used to express a wide variety of intentions. They are also used to fill gaps that arise in conversation when the speaker is pausing to think. Sometimes they are used emphatically to convey the speaker's emotional state. In Japanese discourse, periodic nonverbal utterances called *aizuchi*, are used to indicate that the listener is following what the speaker has said and to create a certain degree of intimacy.

Because nonverbals have minimal lexical structure, they are virtually language-independent. Nonverbals frequently provide concise replacements for words and phrases. The speaker does not have to engage in processes which lead to the formulation of a well-structured phrase or sentence; therefore, the overall response time is significantly reduced. Nonverbals are often comprised of a single phonemic sound, and one such sound could be used to express a number of different emotions and intentions. The listener must interpret the prosodic features of the utterance in order to derive the correct meaning. Distinctions may be understood in terms of the time-variant behavior of the fundamental frequency, amplitude, and formant frequencies. For example, when a speaker utters a nonverbal during a state of confusion, the fundamental frequency will increase with respect to time over most of the utterance duration.

By supplementing speech synthesis systems with the ability to utter nonverbals, discourse could possess a greater degree of naturalness. The computer could be endowed with a flexible and efficient form of expression to add to its repertoire. Emphasis, conveyed through emotion and intention, would be easily perceived through the interpretation of prosodic features. Thus, synthesized speech would not only be more pleasant to the ear but also more meaningful to the mind.

The authors have begun to observe prosodic features of nonverbal expressions, with respect to emotions and intentions, for the purpose of achieving accurate synthesis.

IV. EXPERIMENTS

In order to establish some fundamental relationships among emotions, intentions, and prosody, the authors initiated data collection of utterances including nonverbal expressions.

Four native Japanese speakers, actors by profession, were employed as speakers due to their ability to generate emotion and intention within various situational environments. The speakers, 2 males and 2 females, were requested to utter nonverbal expressions for a variety of emotions and intentions. Utterances were collected for simulated situations, in which the actors improvised an emotional or intentional state, and for natural situations, in which the actors participated in free conversation.

The recorded nonverbals were analyzed with respect to the time-variant behavior of the fundamental frequency, amplitude, and formant frequencies. A selected subset of utterances was used for a perception test in order to determine which emotions and intentions were most easily distinguished. The categories for this subset were *doui*, *kyoumi*, *mukanjou*, *naruhodo*, *odoroki*, *taikutsu*, and *utagai*. English equivalents are listed in Table 1. Thirteen listeners classified the nonverbals according to the emotions and/or intentions they perceived. The listeners were allowed to select from a broader group of categories in order to provide greater freedom of choice. These categories were *doui*, *ikari*, *kanashimi*, *keibetsu*, *kyoumi*, *naruhodo*, *odoroki*, *osore*, *taikutsu*, *tanoshimi*, and *utagai*. (See Table 1).

Japanese	English
doui	consent, agreement
ikari	anger, indignation
kanashimi	sadness, grief
kyoumi	expressing an interest (in)
mukanjou	emotionless, without feeling
naruhodo	expressing "oh I see", "indeed"
odoroki	surprise, astonishment
osore	fear, dread, horror
taikutsu	boredom, dull, being tired (of)
tanoshimi	joy, delight, pleasure
utagai	doubtful, uncertain

Table 1 - List of emotions and intentions used for experiments and their English equivalents

General observations formulated from the perception test are as follows:

1) *utagai* and *odoroki* - utterances collected for the category of *utagai* were almost always classified as both *utagai* and *odoroki*. The overall pitch behavior for these utterances was a rising pattern for most of the duration, followed by a sharp fall at the end as shown in Fig. 1(a). The pitch range was the widest among all of the categories, as much as 370 Hz. Utterances that were classified as belonging only to the *utagai* category were characterized as having narrower pitch ranges and shorter durations compared to those categorized as both *utagai* and *odoroki*.

2) *taikutsu* and *kanashimi* - utterances collected for the category of *taikutsu* were classified as both *taikutsu* and *kanashimi*. The overall pitch behavior was a general declining pattern across the duration of the utterance as shown in Fig. 1(b). The pitch range was the narrowest among all of the categories, approximately 40 Hz. In addition, the duration of these utterances was the longest among all of the categories, sometimes as long as 1700 msec.

3) *doui* and *taikutsu* - utterances collected for the category of *mukanjou* were classified as both *doui* and *taikutsu*. Utterances collected for *doui* were always correctly classified. The overall pitch behavior was a general declining pattern across the duration of the utterance as shown in Fig. 1(c). The pitch range was approximately 70 Hz. The duration for these utterances was the shortest of among all of the categories, with an approximate range of 300 to 600 msec.

4) *naruhodo* and *kyoumi* - utterances collected for these categories were classified as both *naruhodo* and *kyoumi*. The overall pitch behavior was a rising pattern at the beginning and a declining pattern at the end as shown in Fig. 1(d). The pitch behavior for the middle portion was either a fairly stable or a slowly rising behavior. The pitch range seemed to vary according to the intensity of the expression; some were as narrow as 60 Hz but others were as wide as 200 Hz. The duration for these utterances ranged from 600 to 1000 msec.

The data shows that there indeed exist distinguishing prosodic features among some of the categories; however, there is an indication that overlap does exist in some cases.

V. CONSTRAINT PROCESSING

Constraint-based systems have been shown to be successful in a wide variety of applications such as interactive simulations, algorithm animation, and graphical user interface construction [7]. The general philosophy is to treat constraint satisfaction as a minimization problem. The system is defined as having low error when a solution, which best satisfies the constraints, has been found. The system iterates through the constraint satisfaction process until this goal has been attained.

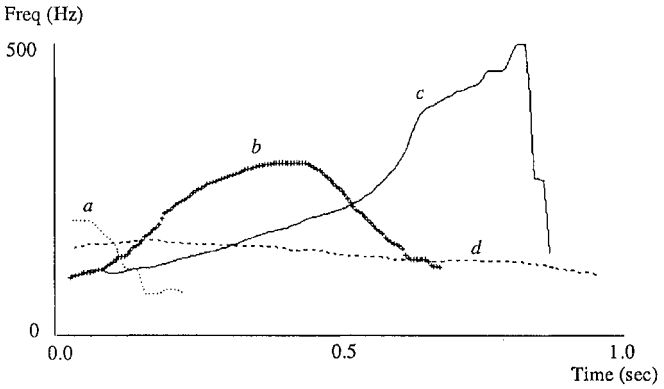


Fig. 1 - Examples time-frequency behavior for nonverbals
(a) doui
(b) kyoumi & naruhodo
(c) odoroki & utagai
(d) taikutsu & kanashimi

The authors favor a constraint-based approach to synthesis in order to avoid the one-to-one mapping limitation of rule-based techniques. However, due to the inherent ambiguities of speech and the desire to use less-mathematically stringent methods in this initial implementation, fuzzy logic is being applied to constraint processing. Fuzzy logic systems have found widespread use in many practical applications because of their ease of use and ability to deal with the ambiguities of real-world events [8].

In contrast to traditional, double-valued Boolean logic, fuzzy logic considers all values from 0 to 1, inclusive. This notion is applied to traditional sets by prescribing a "degree of belongingness" to each element of the set through the declaration of membership functions. The result is a so-called fuzzy set. In this framework, numerical definition of linguistic qualifiers, such as "low" or "high", is possible through the creation of fuzzy sets. A simplified example of a fuzzy set for frequency is shown in Fig. 2.

frequency =
{0,100,200,300,400}
low membership function =
{1.0,0.7,0.4,0.1,0.0}
high membership function =
{0.0,0.1,0.4,0.7,1.0}

where frequency 100 has a level of low of 0.7
and a level of high of 0.1

Fig. 2 - Example of a fuzzy set and its membership functions

For the synthesis of nonverbal expressions, a set of constraints was defined according to the observed data. The constraints were linguistically declared in the form of conditionals; some simple examples are shown in Fig. 3. A fuzzy inference system [9] was used to generate curves for each constrained parameter according to their defined membership functions.

```
If time is early then
        amplitude is soft and frequency is medium.
If time is late then
        amplitude is soft and frequency is very high.
```

Fig. 3 - Examples of linguistically declared constraints

The curves were tuned by adjusting the appropriate membership functions and were input to a formant synthesizer [10] for the generation of nonverbals. Preliminary synthesis results indicate the potential of utilizing fuzzy logic; however, the method for determining membership functions needs to be addressed.

VI. CONCLUSION

As technology moves toward the realization of human-computer interactive systems, researchers should regard naturalness with the utmost importance. The success of a system depends on the user's ability to comfortably utilize system features to their full potential. The user should not have to adapt personal behavior to match the capabilities of the system, rather the system should be adapted to match the personal behavior of the user.

One aspect to be considered is synthesized speech. The most natural form of communication among humans is speech; therefore, its integration into interactive systems is crucial. However, it is not enough to merely synthesize recognizable speech. The speech should embody the animated quality characteristic to that of humans.

A facet of speech which promotes this quality is nonverbal expressions which convey emotions and intentions and add naturalness to discourse. Nonverbal expressions require knowledge representations to hold their compositional information. Because they impose constraints on human speech production processes, they should be defined in terms of constraints within synthesis systems as well. Fuzzy inference provides a simple, yet sufficient, method for maintaining the constraints as it allows for the ambiguities inherent to speech production while providing means for fine tuning.

Synthesis of nonverbals is definitely within reach. Further investigation into prosodic features is necessary for deriving relationships with greater accuracy and specifying constraints with greater precision. Additional parameters need to be constrained in terms of the effects of emotion and intention.

ACKNOWLEDGEMENTS

The authors would like to thank the four actors, Hajime Omori, Mayumi Taki, Mayumi Ushijima, and Shinichi Ushijima for their insight and contributions to the data collection experiments. Kris Maeda is deeply grateful to John Maeda for suggesting the use of fuzzy inference to shape curves, as well as providing inspiration and support.

REFERENCES

[1] K. Hirose, H. Kawai, and H. Fujisaki, "Synthesis of Prosodic Features of Japanese Sentences," Second Symposium on Advanced Man-Machine Interface Through Spoken Language, pp. 1-13, Nov. 1988.

[2] J.P. Cahn, "Generating Expression in Synthesized Speech," Technical Report, Massachusetts Institute of Technology, 1990.

[3] J. Pierrehumbert and J. Hirschberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," in *Intentions in Communication*, MIT Press, Ch. 14, 1990.

[4] B.J. Grosz, "Attentions, Intentions, and the Structure of Discourse," Computational Linguistics, Vol. 12, No. 3, pp. 175-204, Jul.-Sept. 1986.

[5] A. Ortony, G.L. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, 1988.

[6] M. Dyer, "Emotions and their Computations: Three Computer Models," Cognition and Emotion, Vol. 1, No. 3, pp. 323-346, 1987.

[7] A. Borning, R. Duisberg, B. Freeman-Benson, A. Kramer, and M. Woolf, "Constraint Hierarchies," OOPSLA Proceedings, pp. 48-60, Oct. 1987.

[8] L.A. Zadeh, "Fuzzy Sets," Information and Control, Vol. 8, pp. 338-353, 1965.

[9] J.T. Maeda, "A Small Fuzzy Inference System," Technical Report, Nihon University Media Research Laboratory, Oct. 1990.

[10] D.H. Klatt, "Software for a Cascade/Parallel Formant Synthesizer," JASA, Vol. 67, No. 3, pp. 971-995, Mar. 1980.