

Emotional Speech as an Effective Interface for People with Special Needs

Akemi Iida
Keio University, Graduate School of
Media and Governance
5322 Endo, Fujisawa,
Kanagawa, 252-8520 JAPAN
akeiida@sfc.keio.ac.jp

Nick Campbell
ATR Interpreting Telecommunications
Research Laboratories
2-2 Hikaridai, Seika-cho,
Soraku-gun, Kyoto 619-02 JAPAN

Michiaki Yasumura
Keio University, Graduate School of
Media and Governance
5322 Endo, Fujisawa,
Kanagawa, 252-8520 JAPAN
yasumura@sfc.keio.ac.jp

Abstract

This paper describes an application concept of an affective communication system for people with disabilities and elderly people, summarizes the universal nature of emotion and its vocal expression, and reports on the work on designing a corpus database of emotional speech for a speech synthesis in the proposed system.

Three corpora of emotional speech (joy, anger and sadness) have been designed and tested for the use with CHATR, the concatenated speech synthesis system at ATR. Each text corpus was designed to bring out a speaker's emotion. The result of perceptual experiments was proved to be significant and so was the result of CHATR synthesized speech. This indicates that the subjects successfully identified the emotion types of the synthesized speech from implicit phonetic information and hence this study has proved the validity of using a corpus of emotional speech as a database for the concatenated speech synthesis system.

1. Introduction

The authors' ultimate goal is to develop a communication system which can be used by people with disabilities. One of the authors had met people with paralysis arising from cerebral diseases, and had read autobiographies of people suffering from muscular dystro-

phies and cerebral diseases. Most of the patients who suffer cerebral diseases has aphasia along with body paralysis as an aftereffect. She was greatly disturbed to discover that all of them strongly wish to have others understand the existence of their emotion. Even though they cannot speak nor change their facial expressions, they have feelings just as healthy people do and are in need to express them.

Needless to say, people without health problems can also benefit from easy access to the electronic communication. The information technology of the past decades has changed the communication style in Japan significantly to an electronic one. Electronic forums for elderly have been in operation for several years [1] and in the next century, various services will start in the cyberspace: Electronic commerce, tele-shopping, tele-medicine system, video conferencing and so on. Many organizations including welfare offices and NPO's[2] have begun providing information on World Wide Web. Moreover, municipalities plan to issue transcripts such as resident cards electronically within five years.

The preparatory survey conducted by the authors shows that even elderly people without computer skills are interested in electronic services (Subjects were altogether 16; 4 each from 4 groups). They showed their keen interests in tele-medicine system and communication with others (Fig. 1). However, while it might be easy for the younger generation to learn computers, it is not the case with elderly people. Especially

in Japan, where people have not been used to type-writing, using a keyboard as an input device is a great obstacle. Also, deterioration of eyesight is inevitable for the elderly. Therefore, the speech interface both synthesis and recognition would be key techniques for their use.

As Nagasato points out, the coming 21st century will be an aging society. Demographic reports show that the total fertility rate (indicating how many children are born per woman) in advanced countries is less than two, and that in the year 2020, the 1/4 of population in Japan would be over 65 years old[3]. The easy-to-use tool that enables affective communication is desired by the society.

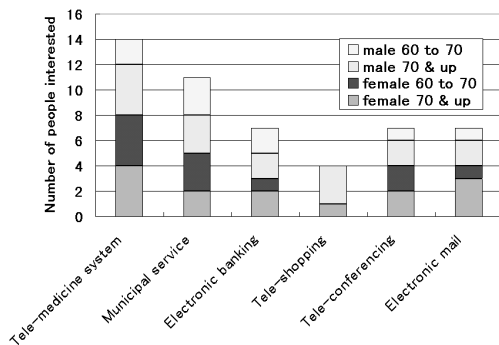


Figure 1. Interests in internet services among the elderly people

With the above motivation, the authors propose an application concept for an affective communication system for people with speaking disabilities and elderly people as illustrated in Fig. 2.

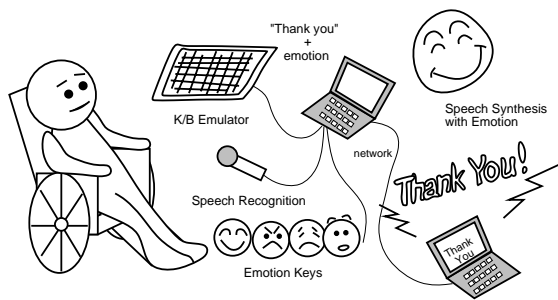


Figure 2. System image

The user can input messages either with a speech recognition system or using a tablet combined with

“emotion keys.” Output will be made using synthesized speech where affective information is reflected. Such system can be extended for severely paralyzed patients where a keyboard emulator is adopted as an input device. Depending on the nature of his/her handicap and preference, the user can connect input devices such as a puff switch, a touch switch or a “one-input mouse” (a device consist of eight arrow buttons to substitute mouse function developed by Ito)[4] to a keyboard emulator. The future image of this system allows us to incorporate speech recognition which can identify a speaker’s emotion.

2. The Nature of Emotion and its Vocal Expression

Emotion plays an important role not only for human lives but also for lives of other animals. According to Frijda and Moffat[5], emotion is described as a change in the state of readiness for maintaining or modifying the relationship with the environment. They proceed by saying that emotion consists of the subjective awareness of those changes and of the events that are relevant to the individual’s concerns which form the basis for the individual’s preferences of states of the world. In other words, it can be said that the function of emotion is to help detect events relevant to the individual’s concerns and, when such an event is detected, to evoke a necessary actions to confront or cope with the event[6].

There are various types of emotion, and categorizing them is a difficult task. Although Shaver and others describes emotional categories as fuzzy sets, they have clustered 135 emotion types to a hierarchical tree structure. The category clustering reaches 6 top nodes which are love, joy, anger, surprise, sadness and fear, though surprise appears to have different characteristics from the other emotion types[7]. The vocal cue is one of the fundamental expressions of emotion, on a par with facial expression. Primates, dolphins, dogs, and all the mammals have emotions and can convey them through vocal cues. People can express emotion by crying, laughing, shouting and also by more subtle characteristics of their speech. Shimura and Imaizumi conducting a perceptual experiment with infants’ vocalization report that infants of two months old can convey emotion by vocalization excluding cries[8]. Although segmental features and the content of the utterance themselves carry emotion, suprasegmental features (such as accent and intonation) play an important part in conveying emotion[9]. Murray and Arnott have conducted a literature review on human vocal emotion and concluded that in general, the acoustic characteristics noted are consistent among different studies carried

Table 1. Acoustic tendencies in previous studies[10]

Speech rate	Fear < Anger < Sadness < Disgust	Happiness cannot be judged
Pitch average	Disgust < Sadness < Happiness < Fear, Anger	
Pitch range	Sadness < Disgust < Fear, Anger, Happiness	
Intensity	Sadness < Disgust < Fear < Anger, Happiness	

Table 2. Acoustic tendencies in authors' study

Duration	Anger < Surprise < Joy < Sadness < Disgust
Pitch range	Sadness < Disgust < Surprise < Anger < Joy
Intensity	Sadness < Disgust < Joy < Surprise < Anger

out, with only minor differences being apparent[10]. The acoustic tendencies of the primary five emotions (anger, happiness, sadness, fear and disgust) are described in their work. Table 1 is the summary of their survey of the most frequently studied acoustic characteristics in researches of vocal emotion.

The acoustic tendency of emotional speech examined above can also be observed in studies in Japan. Various studies has been conducted such as of Kitahara, Shigenaga and Hirose[11, 12, 13]. The relationship between emotional speech and its perceptual impression is described in the authors' previous work[14]. In that experiment, four phrases were spoken by a male and two females with five emotion types (anger, joy, sadness, surprise and disgust). Acoustic parameters measured were duration, intensity and pitch range. SD (Semantic Differential) rating was conducted and the result was analyzed by a principal factor analysis. Factor scores after varimax orthogonal rotation show a correlation between sadness and disgust, which are concentrated in the 'gloomy-weak' impression region, joy and surprise, in 'cheerful' region, and anger in the 'powerful' region (Fig. 3). An experiment where subjects were asked to identify the emotion types was also conducted in that work, and results followed the tendency reported in other researchers' studies (Table 2).

3. Designing and Testing a Corpus of Emotional Speech for Speech Synthesis

As mentioned in the previous section, emotion plays an important role in human communication and the realization of emotion in synthesized speech could lead to many useful applications. Among them, the most urgently required is a communication tool for people

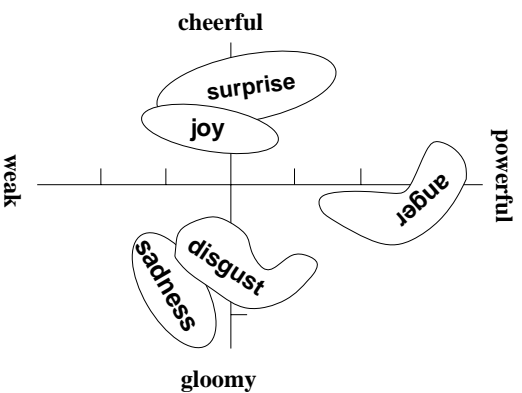


Figure 3. Correlation between emotional voice and perceptual impression

with disabilities.

When the quality of speech synthesis is concerned, speech synthesis until recently was very far from natural-sounding as described in Campbell[15]. According to Campbell, some isolated vowels and consonants could be very well replicated and, with careful hand-turning, even whole utterances could be mimicked, but there was no speech synthesis-by-rule system that could be mistaken for a human speaking[15]. On the other hand, people's demand for natural quality can be easily shown by conducting a simple experiment. The authors gave ten subjects a small task of creating an invitation card by a word processor.

The instructions were given randomly in three different voices (pleasant, sad and angry). Although the difference in subjects' attitude could not be measured quantitatively, the answers for questionnaire showed that all subjects preferred a pleasant voice, and the angry voice made them feel uneasy and depressed.

At present, the best rule-generated synthetic speech that can be heard today is concatenative, using small segments from recorded sequences of real speech and joining them to form novel utterances. The CHATR synthesis system which is being developed by ATR, represents such a system and the author has developed three corpora of emotional speech for use with CHATR. Perceptual experiments were conducted to identify the emotion type of each speech corpus, and of resynthesized speech when using each corpus in turn as a source database. The remaining section of this paper describes the design strategy of the corpus and its acoustic characteristics. It also describes the result and findings of the perceptual experiments.

3.1. Designing a Text Corpus Expressing Emotions

Texts expressing joy, anger, sadness were gathered from newspapers, WWW and self-published autobiographies of disabled people. Since it was desirable for a particular emotion to be sustained for relatively long period of time, monologue texts were selected. Each text corpus is composed as Table 3.

Seventy-two students were asked to judge the emotion category of each text from the combined corpus. All texts but two were identically judged as the emotion types which the corpus designer classified (Fig. 4).

Table 3. Details of text corpus

	Texts	Sentences	Moras	Phonemes
Joy	12	461	21,676	40,916
Anger	15	495	21,085	39,171
Sadness	9	426	16,189	31,840

3.2. Characteristics of a Corpus of Emotional Speech

An adult female read all texts in a sound treated room and the speech was digitized at a 16kHz 16bit sampling rate. Mean fundamental frequency (f_0) of the 'sad' corpus was lower than that for 'anger' and 'joy' (sad: 243Hz, joy: 257Hz, anger: 263Hz) and the standard deviation of the 'sad' corpus was smaller than

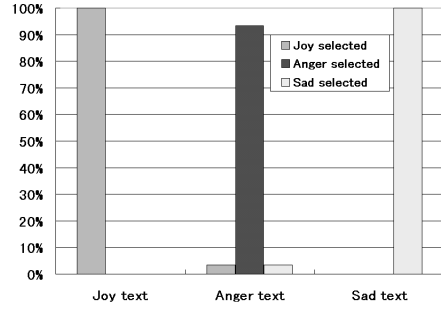


Figure 4. Evaluation of text corpus

those of 'anger' and 'joy' (sad: 40Hz, joy: 53Hz, anger: 57Hz) (Table 4, Fig. 5).

The duration of pauses within a sentence were also measured, and it was found that pauses for the 'sad' corpus were longer than those of 'joy' and 'anger' corpora (Fig. 6).

Table 4. Mean f_0 and its SD

Emotion type	Mean	SD
Joy	256.59	52.90
Angry	262.46	57.26
Sadness	242.91	40.04

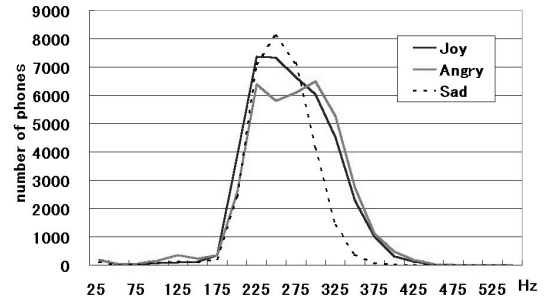


Figure 5. Pitch range per emotion

3.3. Evaluation of a Corpus of Emotional Speech

All sentences in the combined corpus of emotional speech were randomized and presented to twenty-nine

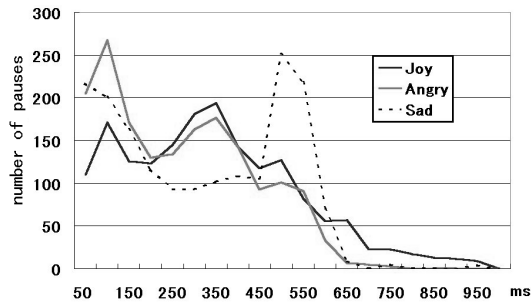


Figure 6. Duration of pauses within a sentence per emotion

university students in order to avoid any contextual interference. Students were asked to select one emotion from joy, anger and sad and as an option, to mark “Cannot be classified as one of the three,” “No intonation nor emotion,” “Can be judged from the context” and “Typical expression for a certain emotion type.” For the latter question, students were allowed to select multiple answers or leave blank any sentences they felt could not be described by the above categories. Result showed joy: 80%, anger: 86%, and sadness: 93% correctly recognized at a significance of $p < 0.01$ (Fig. 7).

3.4. Evaluation of CHATR, Synthesized Speech

Using the corpus of emotional speech which we created as a database, synthesized speech with emotion was developed with CHATR. Eighteen university students were told to identify the emotion types of five context independent synthesized speech produced with source corpora from the three different emotions (joy, anger, sadness). Results showed joy: 51%, anger: 60%, sadness: 82% correctly recognized (at a significance of $p < 0.01$). Chance results can be expected to be around 30%, so we conclude that the characteristics of the emotion are well preserved in the voice (Fig. 8).

3.5. Discussion

Although randomly presented, 47% of sentences in evaluation were marked “Can be judged from the context,” for the source corpus of human emotional speech while only 13% were detected from the context of the CHATR speech synthesis. With the positive re-

sult of the perceptual experiment, this indicates that subjects judged emotion categories not from the explicit context but from the phonetic/acoustic information. Hence, it can be said that authors’ approach of gathering context-dependent texts to bring out typical phonetic information per emotion type is valid. Furthermore, 23% of sentences in evaluation were marked “No intonation nor emotion,” for the corpus of human emotional speech, compared with 27% for CHATR, although the identifying rate was the same with those with no mark for that item. This implies that certain information about emotion is included within the speech tokens themselves.

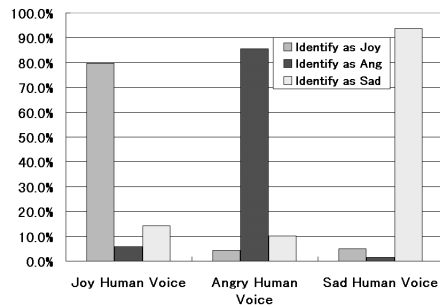


Figure 7. Evaluation of emotional human speech

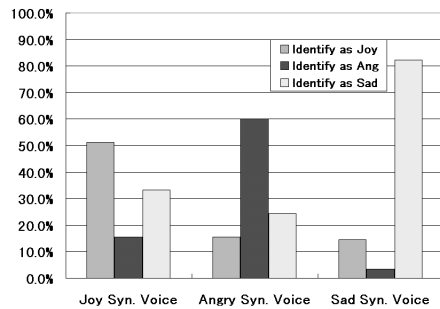


Figure 8. Evaluation of emotional synthesized speech

3.6. Conclusion

A corpus of emotional speech has been created. The text corpus was designed to bring out a speaker’s emotion and to sustain that emotion for a certain period

of time. Each text corpus was read and digitized at a 16kHz, 16bits sampling rate. Acoustic analysis of pitch and duration of pauses showed that the 'sad' corpus can be distinguished from 'anger' and 'joy' corpora. The result of perceptual experiments was proved to be significant and so was the result of synthesized speech. From these results, it can be said that subjects identified the emotion types of the synthesized speech from implicit phonetic information. This study has revealed the validity of using a corpus of emotional speech for synthesis of emotion. Future work will include an analysis of emotion cues in the spectrum and voice source quality for a better quality of synthesis speech and also for the research of emotion recognition in speech.

4. Concluding Remarks

The application concept of an affective communication system for people with speaking disabilities and elderly people has been described. The paper described the universal nature of emotion and its vocal expression. Although various languages are spoken all over the world, emotional expression in speech can be identified beyond the language boundaries. The authors described their work on designing and testing a combined corpus of emotional speech and presented the result which indicated a positive outlook.

Acknowledgement

Authors would like to express their appreciation to Mr. Kazuyuki Ashimura and Ms. Yoko Ohta of ATR-ITL and Dr. Fumito Higuchi, Mr. Soichiro Iga, and Mr. Kentaro Meiseki of Keio University for their support. Authors further would like to thank Prof. Keiichi Hirose and Mr. Hiromichi Kawanami of the University of Tokyo for their kind guidance on speech processing. Special thanks are owe to students at Keio University for participating in their experiments.

References

- [1] <https://iw.nim.niftyserve.or.jp/hns/nifty/fmellow>
- [2] Prop Station, *KSK Flanker* Vol. 17, Apr., 1997.
- [3] Y. Nagasato, T. Takino, H. Yoshikawa. Human-machine interface technology to energize an aging society, In *Human Interface*, vol. 12, pp. 387-394, 1997.
- [4] H. Ito, M. Ohashi, T. Tamagaki, K. Kitamura. A Telecommunication network system which changed severely physically handicapped people's life style and awareness., In *Transactions of Information processing society of Japan*, Vol. 37, No. 5, pp. 931-939 , 1996.
- [5] N. H. Frijda, and D. Moffat. Modeling emotion In *Cognitive studies: Bulletin of the Japanese Cognitive Science Society*, Vol.1, No. 2, 1994.
- [6] M. Toda. *Man, robot and society*, The Hague: Nijhoff, 1981.
- [7] P. Shaver. Emotion Knowledge: Further exploration of a prototype approach, In *Journal of Personality and Social Psychology*, 1987.
- [8] Y. Shimura and S. Imaizumi. Emotional Information in Young Infants' Vocalizations, In *Proc. of ICPHS 1995*, Vol. 3, pp. 412-415, 1995.
- [9] R. Bance and K. Scherer. Acoustic Profiles in Vocal Emotion Expression, In *Journal of Personality and Social Psychology*, 1996.
- [10] I. R. Murray, and J. L. Arnott. Towards the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion, In *Journal of Acoustic Society of America*, Vol. 93, No. 2, pp. 1097-1108, Feb, 1993.
- [11] Y. Kitahara, and Y. Tookura. Prosodic Components of Speech in the Expression of Emotions", In *Proc. ASA/ASJ Joint meeting*, 1989.
- [12] M. Shigenaga. Characteristic Feature of Emotionally Uttered voices revealed by discriminant analysis - on Normalization methods, In *Proc. ASJ Spring Meeting I*, pp. 201-202, 1997.
- [13] K. Hirose, N. Takahashi, H. Fujisaki, S. Ohno. Representation of Intention and Emotion of Speakers with Fundamental Frequency Contours of Speech, In *Technical Report of the Institute of Electronics, Information and Communication Engineers*, HC94-41, pp.33-40, Sept, 1994.
- [14] A. Iida, S. Iga, M. Yasumura. Towards the realization of speech synthesis with emotion, In *HCI international 97 Poster Session Proceedings*, p.41, 1997.
- [15] W. N. Campbell. From Read Speech to Real Speech, In *ICPhS95 Stockholm*, Vol. 2, pp. 20-27, 1995.
- [16] W. N. Campbell. Chatr: a high-definition speech synthesis system, In *Proc. ASA/ASJ Joint meeting*, 1996.