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 Hikaridaie, 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 effective communication system for people with disabilities and elderly people, summarizes the universal nature of emotion and its vocal expression, and reports on designing a corpus database of emotional speech for speech synthesis system, a key technology to be adopted by the proposed system.

We have designed and tested three corpora of emotional speech (joy, anger and sadness) for the use with CHATR, the concatenated speech synthesis system at ATR. Each text corpus was designed to bring out a speaker’s emotion. Acoustic analysis of pitch and duration of pauses showed that the ‘sad’ corpus can be distinguished from ‘anger’ and ‘joy’. 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 emotive 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 dystrophies and cerebral diseases. Most of the patients who suffer cerebral diseases suffer 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 each group). They showed their keen interests in Tele-medicine system and communi-
cation 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 typewriting, 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 is 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.

As shown in Fig.1, there is an increasing interest 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.

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 paralysed 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 Prijda 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 action 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 Imaiimi conducting a perceptual experiment with infants' vocalization[8] report that infants of two months old can convey emotion by vocalization excluding cries. Although segmental features and the content of the utterance themselves carry emotion, suprasegmental features (such as accent and intonation) play an im-
portant part in conveying emotion[9]. Murray and
Arnott have conducted a literature review on human
vocal emotion and concluded that in general, the acous-
tic characteristics noted are consistent among different
studies carried out, with only minor differences being
apparent[10]. The acoustic tendencies of the primary
five emotions (anger, happiness, sadness, fear and dis-
gust) are described in their work. Table 1 is the sum-
mary 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,
Shigenaka and Hirose[11, 12, 13]. The relationship
between emotional speech and its perceptual im-
pression 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 para-
eters 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 sub-
jects were asked to identify the emotion types was also
conducted in that work, and results followed the ten-
dency 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
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]. Ac-
cording to Campbell, some isolated vowels and conso-
nants could be very well replicated and, with careful
hand-turning, even whole utterances could be mim-
icked, but there was no speech synthesis-by-rule sys-

tem that could be mistaken for a human speaking[15].

On the other hand, people’s demand for natural qual-
ity can be easily shown by conducting a simple ex-
periment. The authors gave ten subjects a small task
of creating an invitation card by a word processor.

![Figure 3. Correlation between emotional voice
and perceptual impression](image)

The instructions were given randomly in three different
voices (pleasant, sad and angry). Although the differ-
ence in subjects’ attitude could not be measured quan-
titatively, 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 de-
veloped 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 reports on 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 autobi-
ographies 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.
Table 1. Acoustic Tendencies in previous studies

<table>
<thead>
<tr>
<th>Speech rate</th>
<th>Fear &lt; Anger &lt; Sadness &lt; Disgust &lt; Happiness: cannot tell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch average</td>
<td>Disgust &lt; Sadness &lt; Happiness &lt; Fear, Anger</td>
</tr>
<tr>
<td>Pitch range</td>
<td>Sadness &lt; Disgust &lt; Fear, Anger, Happiness</td>
</tr>
<tr>
<td>Intensity</td>
<td>Sadness &lt; Disgust &lt; Fear &lt; Anger, Happiness</td>
</tr>
</tbody>
</table>

Table 2. Acoustic Tendencies

<table>
<thead>
<tr>
<th>Duration</th>
<th>Anger &lt; Surprise &lt; Joy &lt; Sadness &lt; Disgust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity</td>
<td>Sadness &lt; Disgust &lt; Joy &lt; Surprise &lt; Anger</td>
</tr>
<tr>
<td>Pitch</td>
<td>Sadness &lt; Disgust &lt; Surprise &lt; Anger &lt; Joy</td>
</tr>
</tbody>
</table>

Table 3. Details of text corpus

<table>
<thead>
<tr>
<th>Texts</th>
<th>Sentences</th>
<th>Morae</th>
<th>Phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joy</td>
<td>12</td>
<td>461</td>
<td>21,576</td>
</tr>
<tr>
<td>Anger</td>
<td>15</td>
<td>495</td>
<td>21,085</td>
</tr>
<tr>
<td>Sadness</td>
<td>9</td>
<td>426</td>
<td>16,189</td>
</tr>
</tbody>
</table>

72 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).

Figure 4. Evaluation of text corpus

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

Figure 5. Pitch range per emotion

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).

3.3. Evaluation of a Corpus of Emotional Speech

All sentences in the combined corpus of emotional speech were randomized and presented to 29 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
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. 18 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 result 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.

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 reported on their on work 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 Ashinura and Ms. Yoko Ohta of ATR-ITL and Mr. Kentaro Melsen of Keio University for their support. Authors further would like to thank Dr. Keikichi Hirose and Mr. Hiromichi Kawanami of the University of Tokyo for their kind guidance on speech processing. Special thanks are owed to students at Keio University for participating in our experiments.

References

[1] https://iw.nim.niftyserve.or.jp/hns/nifty/fmellow