ACOUSTIC NATURE AND PERCEPTUAL TESTING OF CORPORA OF EMOTIONAL SPEECH

Akemi Iida, Nick Campbell**, Soichiro Iga*, Fumito Higuchi*, Michiaki Yasumura**

**Graduate School of Media and Governance, Keio University,*
5322, Endo, Fujisawa, Kanagawa, 252-8520, Japan
***ATR Interpreting Telecommunications Research Laboratories*
2-2 Hikaridai, Seika-cho, Kyoto 619-02, Japan
e-mail: akeiida@sfc.keio.ac.jp, nick@itl.atr.co.jp

ABSTRACT

This paper proposes three corpora of emotional speech in Japanese that maximize the expression of each emotion (expressing joy, anger, and sadness) for use with CHATR, the concatenative speech synthesis system being developed at ATR. A perceptual experiment was conducted using the synthesized speech generated from each emotion corpus and the results proved to be significantly identifiable. Authors' current work is to identify the local acoustic features relevant for specifying a particular emotion type. F0 and duration showed significant differences among emotion types. AV (amplitude of voicing source) and GN (glottal noise) also showed differences. This paper reports on the corpus design, the perceptual experiment, and the results of the acoustic analysis.

1. INTRODUCTION

Emotion plays an important role in communication, and vocal expression is one of the fundamental expressions of emotion, on a par with facial expression. The realization of speech synthesis with emotion is a difficult task but it can lead to many useful applications such as communication tools for people with speaking disabilities. Developing a speech corpus for CHATR and having thereby CHATR synthesize emotional speech is a step in this direction.

1.1. A Natural Speech Re-sequencing

Synthetic speech close to natural sounding that can be heard today is concatenative. The CHATR synthesis system of ATR generates such speech. Being a re-sequencing speech synthesizer, CHATR produces an index for a random-access retrieval of waveform sequences from the externally stored corpus to select target units to create new utterances. In so doing, it removes the necessity for signal processing but instead requires a larger library for a source unit [1]. When synthesizing with CHATR, the ideal size of the source database is yet to be explored but preferably 30,000 to 50,000 in phoneme level. This size appears to cover the phonemic and prosodic variations encountered in Japanese.

Three steps are taken for indexing. Converting an orthographic transcription to a phonemic representation of the speech, aligning the phoneme to the waveform to provide a key to the prosodic feature extraction, and producing feature vectors for each phone (identifying label, access information, cepstral distance, f0, duration and power). Each phone holds the feature information of the current, the previous and the following phone. After indexing, CHATR calculates the weight of each feature per phone to determine optimal selection of a unit. A unit is selected by way of maximizing continuity and minimizing the target distance. Two functions are used: "Target cost" and "concatenation cost." The former is an estimate of the difference between a database unit and the latter is an estimate of the quality of a join between consecutive units.

1. 2. Emotion and its Vocal Expression

Emotion is described as a change in the state of readiness for maintaining or modifying the relationships with the environment. There are various types of emotion, and categorizing them is a difficult task. Human can express their feeling by crying, laughing, shouting and also by more subtle characteristics of their speech. In their literature review, Murray and Arnott state that in general, the acoustic characteristics are consistent among different studies carried out, with minor differences being apparent. The tendencies of acoustic features such as f0, power and speech rate of the primary five emotions (anger, happiness, sadness, fear and disgust) are summarized in their work [2]. Voice quality, pitch changes and articulation are also reviewed. The acoustic tendency of emotional speech examined above can also be observed in studies in Japan such as Kitahara [3] and Hirose and others [4]. The relationship between emotional speech and perceptual impressions is described in Iida and others [5].

2. DESIGNING AND TESTING A CORPUS

This paper proposes three corpora of emotional speech in Japanese that maximize the expression of each emotion (expressing joy, anger, and sadness) when read. Perceptual experiments were conducted to identify the emotion type of each speech corpus, and of the resynthesized speech from CHATR when using each corpus in turn as a source database. For both experiments, results proved to be significant.

2.1 Designing a Corpus of Emotions

Although there were various kinds of emotions, 'joy,' 'sadness' and 'anger' were chosen as a first trial. These emotion types appear to be the fundamental emotions that at least people with speaking disability might wish to express [5]. When emotional variation is taken into account, the main text-level requirement for a corpus is to be able to induce natural emotion in the speaker when read. To include proper linguistic and semantic contents is essential, but to be able to induce a natural expression of emotion rather than to simply require an acted or simulated emotion is preferred. We gathered texts expressing joy, anger, and sadness from newspapers, the WWW and self-published autobiographies of disabled people. Monologue texts were chosen so that a particular emotion could be sustained for a relatively long period of time. Some expressions typical to each emotion were inserted in appropriate place in the text in order to maximize the expression of each target emotion [6]. To meet the size requirement for CHATR database, more than 30,000 phonemes were collected for each corpus (Table 1).

	Texts	Sentences	Moras	Phonemes
Joy	12	461	21676	40916
Anger	15	495	21085	39171
Sadness	9	426	16189	31840

Table 1: Size of Each Text Corpus

Although it has become a standard to use identical texts in studies of emotional speech, priority in this study was given to the naturalness of the speech. As a result, three corpora contain completely different texts. The authors knew that this would not present a problem, since the main objectives of these corpora were to serve as a source database for CHATR, where the basic unit for use was not the text but the phone-sized waveform segment. Fig. 1 is an example from 'joy' corpus.

> Joy Text No. 5
> Mattaku teashi no ugokanai watashi nimo jibun de yareru kotoga dekita no desu. "Sugoizo, sugoizo! Oi, konna koto mo dekiruzo! Miteroyo, iika? Hora, mou ichido yattemirukarana! Iya, gokigendayo, kore!"
> (Even I, whose body is completely paralyzed, could do it. "It's great! Just great! Hey, I can do things like this, too! Look at me, are you ready? See, I'll do it again. Oh, it's absolutely fantastic!")

Figure 1: Example text from 'joy' corpus

The phonetic balance of each corpus was not considered for this trial, again due to giving priority to naturalness. If we try to balance the number of phonemes in each corpus, it would not be possible to gather texts from natural sources. We also believe that the phoneme combinations in a sufficient amount of gathered materials are adequate to represent the combinations that would also appear in natural utterances under each emotion type. In fact, our assumption worked as expected with the CHATR's unit selection rule of maximizing continuity using concatenation costs when emotional speech was synthesized.

Neither actors nor actress were used for recording in order to avoid exaggerated expression. An adult female (the first author) read all texts in a sound treated room where good recording level was maintained and the speech was digitized at a 16kHz 16bit sampling rate.

To summarize, our design policy was as follows:

1. Develop only the basic emotions for initial trials.
2. Gather texts written to express natural emotion.
3. Target size of each corpus is 30,000 phonemes.
4. Include typical phrases of the target emotion.
5. Avoid exaggerated expression.
6. No specific consideration for the phonetic balance needed at this size.

2.2 Testing a Corpus of Emotional Speech

Evaluation of the Corpus in Text Level
72 student volunteers were asked to judge the emotion category of each component text from the combined corpus. All texts but two were correctly judged as representing the emotion types the corpus designer classified. This confirms that the content of the passages is sufficiently emotion-rich to be unambiguous. The remaining question now is whether there are cues in speech that also help to distinguish the component emotions.

Evaluation of a Corpus of Emotional Speech
We performed an evaluation of whether emotion type could be recognized from the speech. In order to avoid any contextual interference, all sentences in three corpora were randomized and presented to 29 university students for an emotion-type classification [7].

Since it was impossible to separate the content of an utterance from the style of its speech, we gave subjects a two-part task. The purpose of this was to determine the degree to which emotion could be recognized from the wording of the utterance and the degree to which it was recognizable from the speech. Students were first given a forced-choice selection of joy, anger and sadness for each. In addition, as an option, the following riders were given: "Cannot be classified as any of the three," "No marked emotion," "Can be judged from the textual content" and "Typical expression for a certain emotion type." For these optional riders, students were allowed to select multiple answers or to leave blanks for 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. 2).

Evaluation of CHATR, Synthesized Speech
Using three corpora of emotional speech, we created a source database for emotional synthesizing speech using CHATR. To test the validity of this system, 18 university students were asked to identify the emotion types of five synthesized utterances produced with each source corpora from the three different emotions (total of 15 sentences). To create the sentences, we chose semantically neutral texts, unmarked for emotion such as Fig. 3.

> Chataa wa iroiro na koe de shaberu kotono dekiru atarashii onsei gosei no shisutemu desu. (CHATR is a new speech synthesizer that can speak in various voices).

Figure 3: Example text for synthesis.

Utterances were then synthesized for each text using speech segments selected from each of the three different emotion databases [7]. Results showed joy: 51%, anger: 60%, sadness: 82% correctly recognized at a significance of p 0.01(Fig. 4). Chance results can be expected to be around 30%, so we conclude that the characteristics of the emotion are well preserved in the speech.

Discussion

Although randomly presented, 47% of sentences in the corpus evaluation were marked "Can be judged from the text content," for the source corpus of human emotional speech while only 13% were similarly recognizable from the text in the sentences used for the CHATR speech synthesis. Furthermore, 23% of sentences in the evaluation were marked "No emotion," for the corpus of human emotional speech, compared with 27% for CHATR, although the identifying rate was the same as those with no mark for that rider. Along with the significant scores in emotion identification, this indicated that subjects judged emotion categories not from the explicit content of the individual utterances, but from the phonetic information in the speech and that certain information about emotion was included in the speech units (i.e. phonemes) themselves.

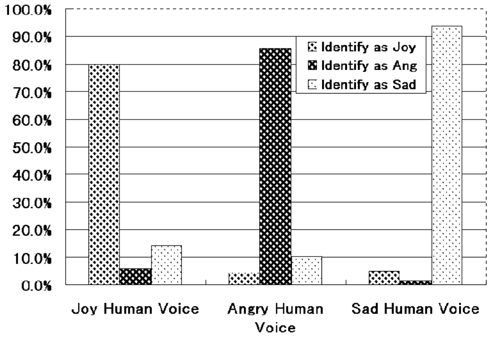

Figure 2: Test Result of Emotional Human Speech

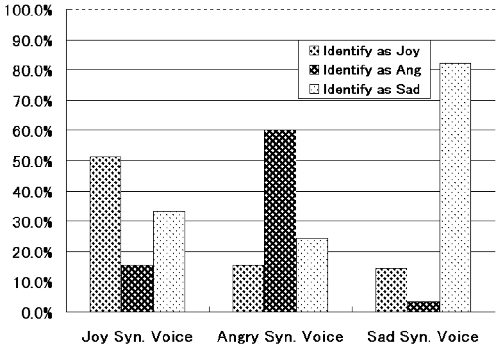

Figure 4: Test Result of Emotional Synthesized Speech

3. ACOUSTICS OF THE CORPUS

We analyzed duration, power, formants and glottal parameters. Here, our objective was to seek relevant features for specifying a particular emotion and if any, to specify in a parametric way.

f0, Duration and Power

Means and standard deviations (SD) of f0, duration and power per phone for each corpus were measured (Table 2). Mean fundamental frequency (f0) of the 'sad' corpus was lower, and SD was smaller than those of 'anger' and 'joy' were. Duration per phone for 'sadness' was the longest and that of 'anger' was the shortest. Means Comparisons by ANOVA showed that for f0, all three types of emotions were significantly different from one another, for duration, 'anger' and 'joy' appeared significantly different than 'sad,' and, for power, no difference.

	f0		Duration		RMS	
	Mean	SD	Mean	SD	Mean	SD
Joy	255.2	52.6	64.8	31.4	6.83	0.63
Ang	260.1	56.9	66.1	28.6	6.77	0.59
Sad	240.8	38.2	73.4	31.8	6.82	0.63

Table 2: Mean and SD of f0, Duration and Power

The duration of pauses within each sentence were also measured, and it was found that pauses for the 'sad' corpus were longer than those of 'joy' and 'anger' corpora (Fig. 5).

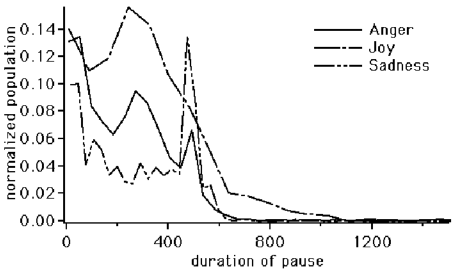

Figure 5: Duration of Pause per Emotion

We investigated the mean of f0, intensity and duration of the vowels where characteristics above were maintained for f0 and duration. Means of f0 and duration are shown in Fig. 6.

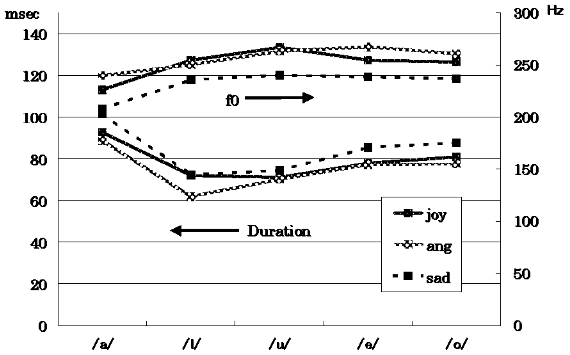

Figure 6: Means of f0 and Duration of Vowels

Formants

Formants were analyzed using ESPS (Entropic Research Lab. Inc) and ARX (Auto-Regressive with Exogenous Input) analysis system [8]. Glottal parameter was analyzed by ARX. Fig. 7 shows not the absolute frequency but relative frequencies of means of entire vowels per emotion taking that of 'angry'

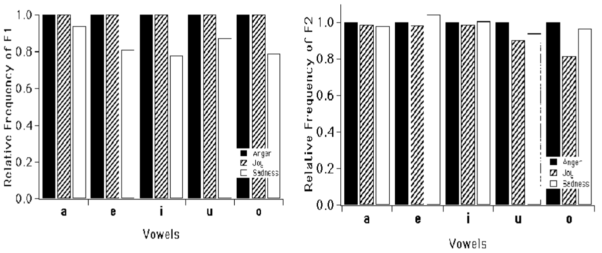

Figure 7: Mean F1, F2 of Vowels per Emotion

vowels as 1.0. F1 and F2 of ‘sad’ vowels were lower than those in ‘anger’ and ‘joy’ were and there seemed to be a correlation between formants and f0. Further, formants of synthesized /a/ extracted from the most correctly identified CHATR speech (The sentence in Fig. 1, Identifying rate = 77%) were analyzed. Mean F0 was in the same order as Table 2 (a>j>s) although formants for this particular vowel did not show correlation with f0. Analyzing formants in natural speech is a difficult task. Vowels in the emotional corpus were very short and formant undershoot might have taken place since most of them were occupied by transitions effected by neighboring consonants.

Glottal Parameters

Glottal parameters were analyzed with ARX for the followings: 1) “Sankaku,” of human speech which were in all three emotional corpora, 2) CHATR synthesized speech shown in Fig. 3, 3) another CHATR speech with second highest identifying score of 74%, “Ah, tsukareta (Oh I’m tired).” Measured parameters were AV (amplitude of voicing source), OQ (open quotient), TL (glottal tilt), STL (spectral tilt) and GN (glottal noise) and f0. Means for all emotion type data were compared by ANOVA. The result showed that there were no significant difference in OQ, TL and STL but there were differences in AV, GN, and f0. Further analysis in glottal parameters is encouraged focusing on AV, GN in addition to f0.

	AV	GN	f0
Human	j>a>s	a>s>j	a>j>s
CHATR1	j>a || >s	a> || s>j	a> || s>j
CHATR2	j>a || >s	j>a || >s	s>j || >s

Table 3: Value Order per Emotion. For all features for Human voice, three types of emotion are significantly different from one another. ‘||’ partitions two groups where the former is significantly different than the later.

Attempt to Minimize the f0 Emotional Cue

Two kinds of modulated speech were synthesized with CHATR to limit the effect of f0 emotional cue as much as possible [7]. 1) Set target f0 to 220Hz, and 2) Set target f0 of ‘joy’ and ‘sad’ to that of ‘anger.’ Although target f0 was set, CHATR got the nearest f0 and as a result, the possibility of unnatural accentuation could not be avoided. Five subjects compared the differences for both sets. For set 1), emotional differences were identified. For set 2), subjects could tell there were differences but could not judge if it was derived from differences in emotion types or from unnatural accentuation. This trial indicates that although f0 is a powerful cue to emotion, other features also serve as emotion cues.

4. CONCLUSION

Context-independent speech corpora of three types of emotion have been developed. A perceptual experiment was conducted using the synthesized speech generated from each emotion corpus and the results proved to be significantly identifiable. Acoustic analysis indicated f0 and duration served as powerful features to cue emotion. Formants, AV and GN also showed differences among three types of emotion. Further effort to identify features that are relevant for specifying a particular emotion in a parametric way is still required in order for those features to be included as selection criteria for CHATR, thereby allowing us to combine the three databases into one corpus which results in smaller sized database. Also, in each corpus, there are speech segments which are less marked for any particular emotion and if those could be identified, they could be used for synthesis of general speech. It has yet to be confirmed that the correlation we found between emotion types and the formants/glottal parameters is sufficient to discriminate between emotion types in the combined corpora.

5. ACKNOWLEDEGMENTS

Authors would like to express their appreciation to Dr. Wen Ding of ATR-ITL for providing ARX analysis system, to Ms. Yoko Ohta of ATR-ITL for preparing the data with CHATR.

6. REFERENCES

1. Campbell, W. N., and Black, A.W., “Chatr: a multi-lingual speech re-sequencing synthesis system,” *Tech. Rept. IEICE SP96-7,* 45-52, 1996.

2. I.R. Murray, I.R. and Arnott, J. L., “Towards the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion,” *J Accost. Soc. Am. 93, No.2*, 1097-1108, 1993.

3. Kitahara, Y. and Tookura, Y., “Prosodic Components of Speech in the Expression of Emotions,” *ASA-ASJ joint meeting*, 1998.

4. Hirose, K., Takahashi, N., Fujisaki, H., Ohno, S., “Representation of Intonation and Emotion of Speakers with Fundamental Frequency Contours of Speech,” *Tech. Rept. Of the IEICE HC94-41*, 33-40, 1994 (Japanese).

5. Iida, A., Campbell, W. N., Yasumura, M., “Designing and testing a corpus of emotional speech,” Proc. *of First International Workshop on East-Asian Language Resources and Evaluation - Oriental Cocosda Workshop '98*, 32-37, 1998.

6. National Language Research Institute, *Bunrui Goi Hyou*, Dainippon Printings, 1964 (Japanese).

7. URL: http://www.sfc.keio.ac.jp/~akeiida/emotion_voice

8. Ding, W. and Campbell, W. N., “On the correlation of prominence and voice source,” *J. Acoust. Soc. Jpn, Proc. of fall meeting*, 197-198, 1996 (Japanese).