ACOUSTIC NATURE AND PERCEPTUAL TESTING OF A CORPUS 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 International Telecommunication Research Laboratories*
2-2 Hikaridai, Seika-cho, Kyoto 619-02, Japan
e-mail: akeiida@sfc.keio.ac.jp, nick@itl.atr.co.jp

ABSTRACT

We have developed three corpora of emotional speech which were designed to 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. Also, AV (amplitude of voicing source) and GN (glottal noise) also showed differences. This paper describes the corpus design, the perceptual experiment, and reports on 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 and create new utterances. In so doing, it removes the necessity for signal processing but instead requires a larger library of source units [1].

Three steps are taken for indexing. Conversion of orthographic transcription to a phonemic representation of the speech, alignment of the phonemes to the waveform to provide a key to the prosodic feature extraction, and production of a feature vector for each phone (identifying label, access information, f0, duration and power). After indexing, CHATR calculates the weight of each feature per phone to determine optimal selection of a unit. CHATR holds cepstral distance, f0, duration and power information of the current, previous and following segments at each phone. The Selection of a unit is done to maximize continuity and minimize the target distance. Two functions are used: "Target cost" is an estimate of the difference between a database unit and "concatenation cost" 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. The vocal cue is one of the fundamental expressions of emotion, on a par with facial expression. Human can express their feeling by crying, laughing, shouting and also by more subtle characteristics of their speech. Murray and Arnott have conducted a literature review on human vocal emotion and states that in general, the acoustic characteristics noted 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 its perceptual impression is described in Iida and others [5].

2. DESIGNING AND TESTING A CORPUS

We have developed three corpora of emotional speech (expressing joy, anger, and sadness) which were designed to maximize the expression of each emotion 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

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.

Although there were various kinds of emotions, joy, sadness and anger were chosen as a first trial since these 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, therefore 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 relatively long periods 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. Structure of Text Corpus

Although it is becoming a standard to use identical texts instudies of emotional voice, priority was given to the naturalness of the speech and as a result, all 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 is not the text but the phone-sized waveform segment. Below is an example from Joy corpus.

> 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!")

Table 2: 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.

Neither actors nor actress were used for recording 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, the following is the our design policy:

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 which 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 clues from the voice quality of emotional 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 the combined original corpus of emotional speech were randomized and presented to 29 university students for an emotion-type classification.

Since it was impossible to separate the content of an utterance from the style of its speech, we gave respondents a two-part task. The purpose of this is to determine the degree to which emotion could be recognized from the wording of the utterance and the degree to which it is recognizable from the voice. Students were first given a forced-choice selection of joy, anger and sadness for each. In addition, as an option, any of 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. 1).

Evaluation of CHATR, Synthesized Speech

Using the three corpora of emotional speech, we created a source database or emotional synthesizing speech using CHATR. To test the validity of this system, 18 university students were asked to identify the emotion types of five context-independent (i.e., text-neutral) synthesized utterances produced with source corpora from the three different emotions (joy, anger, sadness). To create the sentences, we chose semantically neutral texts, unmarked for emotion, such as "Chataa wa iroiro na koe de shaberu kotono dekiru atarashii onsei gosei no shisutemu desu. (CHATR is a new speech synthesizer which can speak in various voices)." Utterances were then synthesized for each text using speech segments selected from each of the three different emotion databases. Results showed joy: 51%, anger: 60%, sadness: 82% correctly recognized at a significance of p 0.01(Fig. 5). Chance results can be expected to be around 30%. so we conclude that the characteristics of the emotion are well preserved in the speech.

Discusstion

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 item. Along with the significant scores in emotion identification, this indicates 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 is included in the speech units (i.e. phonemes) themselves.

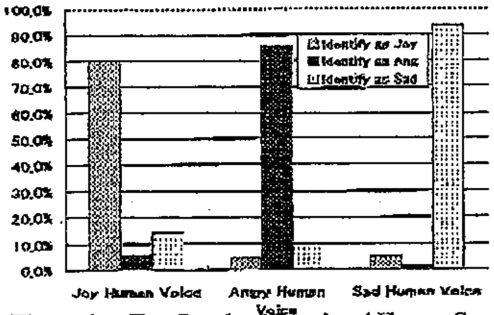

Figure 1: Test Result of Emotional Human Speech

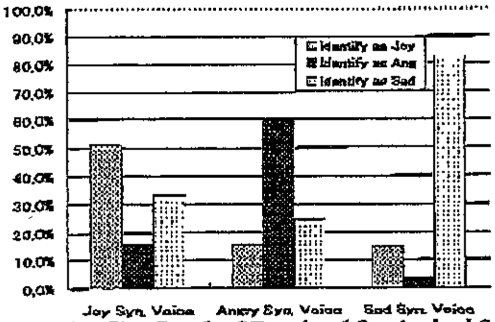

Figure 2: Test Result of Emotional Synthesized Speech

3. ACOUSTICS OF THE CORPUS

We analyzed duration, power, formants and glottal parameters. Here, our objective is 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 5). Mean fundamental frequency (f0) of the 'sad' corpus was lower, and SD was smaller than those of 'anger' and 'joy' (Fig 6). Duration per phone for 'sadness' was the longest and that of 'anger' was the shortest. Means Comparisons by ANOVA showed that all three types of emotions were significantly different from one another for f0. 'anger' and 'joy' appeared significantly different than 'sad' for duration, and, no difference for power.

	f0		Duration		RMS	
	Mean	SD	Mean	SD	Mean	SD
Joy	255.2	52.6	64.8	31.4	6.83	0.63
Anger	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 consistently longer than those of 'joy' and 'anger' corpora.

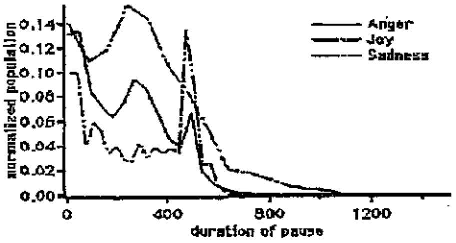

Figure 3: Duration of Pause per Emotion

We investigated the mean of f0, intensity and duration of the vowels where characteristics above are maintained for f0 and duration. Means of f0 and duration are shown in Fig. 4.

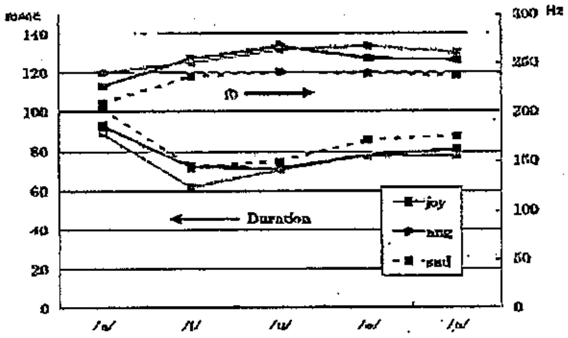

Figure 4: Means of f0 and Duration of Vowels

Formants

Formants were analyzed using ESPS (Entropic Research Lab. Inc) and ARX (Auto-Regressive with Exogenous Input) analysis system [7]. Glottal parameter analysis was done by ARX. Fig. 9 to 11 shows not the absolute frequency but a relative frequency of means of entire vowels per emotion taking that of angry vowels as 1.0. As can be seen in Fig. 9 and 10, f1 and f2 of 'sad' vowels were lower than for the other emotions in

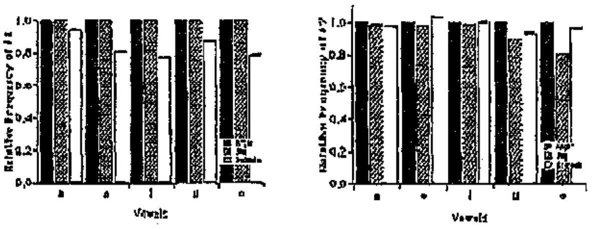

Figure 5: Mean F1, F2 of Vowels per Emotion

most of the vowels and there seems to be some correlation between formants and f0. Further, formants of synthesized /a/ extracted from the most correctly identified CHATR sentence (The sentence shown in previous section of which identified rate was 77%) were analyzed. F0 is in the same order as the above data 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 most of them are occupied by transition effected by the neighboring consonants and as a consequence, formant undershoot might have taken place.

Glottal Parameters

Glottal parameters were analyzed with ARX for the followings. "Sankaku", of human speech which were in all three emotional corpora, CHATR synthesized sentence mentioned above, another CHATR sentence, "Ah, tsukareta (Oh I'm tired)," with second highest identify score of 74% 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 no significant difference in OQ, TL and STL but there were differences in AV, GN, and f0. From this result, further analysis in glottal parameters are encouraged paying attention to AV, GN in addition to f0.can be concentrated on three features.

	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 [10]. 1) Set target f0 to 220Hz, and 2) Set target f0 of 'joy' and 'sad' to that of 'anger.' Although target f0 is set, CHATR gets the nearest f0 and as a result, the possibility of unnatural concatenation cannot be avoided. 5 subjects compared the differences for both sets and for set 1), emotional differences were identified. For set 2), subjects could tell there were differences but could not judge if it is derived from emotional difference or unnatural concatenation. From this trial, we can infer that although f0 is a powerful cue to emotion, other features also serves as emotion cues.

CONCLUSION

A context-independent corpus of three different types of emotional speech has been created and its validity for synthesis of emotional speech was confirmed through perceptual experiments. Acoustic analysis indicated f0 and duration serves as a powerful feature to cue emotion, and formants, AV and GN also showed differences among three types of emotion. Further effort to specify features are relevant for specifying a particular emotion in a parametric way is encouraged so that those features could be included as selection criteria for CHATR, 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, those could be used for synthesis of general speech. The correlation between emotion types and formants/glottal parameters is yet to be studied in depth along with analysis method.

ACKNOWLEDEGMENTS

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

5. 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 fall*, 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 (in 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 (in Japanese).

7. 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 (in Japanese).

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