Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Speech Communication 40 (2003) 161–187

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A corpus-based speech synthesis system with emotion

Akemi Iida [a,b,*], Nick Campbell [c,b], Fumito Higuchi [d], Michiaki Yasumura [d]

[a] *Keio Research Institute at SFC, Keio University, 5322, Endo, Fujisawa-city, Kanagawa, 252-8520, Japan*
[b] *JST (Japan Science and Technology), CREST, Kyoto, Japan*
[c] *ATR Human Information Sciences Research Laboratories, Kyoto, Japan*
[d] *Graduate School of Media & Governance, Keio University, Kanagawa, Japan*

Abstract

We propose a new approach to synthesizing emotional speech by a corpus-based concatenative speech synthesis system (ATR CHATR) using speech corpora of emotional speech. In this study, neither emotional-dependent prosody prediction nor signal processing per se is performed for emotional speech. Instead, a large speech corpus is created per emotion to synthesize speech with the appropriate emotion by simple switching between the emotional corpora. This is made possible by the normalization procedure incorporated in CHATR that transforms its standard predicted prosody range according to the source database in use. We evaluate our approach by creating three kinds of emotional speech corpus (anger, joy, and sadness) from recordings of a male and a female speaker of Japanese. The acoustic characteristics of each corpus are different and the emotions identifiable. The acoustic characteristics of each emotional utterance synthesized by our method show clear correlations to those of each corpus. Perceptual experiments using synthesized speech confirmed that our method can synthesize recognizably emotional speech. We further evaluated the method's intelligibility and the overall impression it gives to the listeners. The results show that the proposed method can synthesize speech with a high intelligibility and gives a favorable impression. With these encouraging results, we have developed a workable text-to-speech system with emotion to support the immediate needs of nonspeaking individuals. This paper describes the proposed method, the design and acoustic characteristics of the corpora, and the results of the perceptual evaluations.

Zusammenfassung

Wir schlagen einen neuen Ansatz vor zur Synthese emotionaler Sprache durch ein verkettetes Sprach-Synthese-System (ATR CHATR) auf Corpus-Basis unter Verwendung von Sprach-Corpora emotionaler Sprache. Mit unserer Methode kann durch einfaches Umschalten zwischen aus dem Corpus generierten Quelldatenbanken Sprache mit angepasster Emotion synthetisiert werden. Wir überprüfen unseren Ansatz durch Generierung dreier Arten emotionaler Sprach-Corpora (Zorn, Freude und Traurigkeit) aus Aufnahmen eines männlichen und eines weiblichen japanischen Sprechers. Die akustischen Eigenschaften jedes Corpus sind unterschiedlich und emotional identifizierbar. Die akustischen Eigenschaften der jeweiligen emotionalen Sprache, die mit unserer Methode synthetisiert wurde, zeigen eine klare Korrelation mit den akustischen Eigenschaften des jeweiligen Corpus. Erkennungsversuche unter Verwendung synthetisierter Sprache zeigen, dass mit unserer Methode emotionale Sprache wiedererkennbar synthetisiert werden kann. Des weiteren überprüfen wir bei unserer Methode die Wortverständlichkeit und den allgemeinen Eindruck, den diese Synthese-Sprache auf den Hörer macht. Die Ergebnisse zeigen, dass durch die vorgeschlagene Methode Sprache

* Corresponding author.

0167-6393/02/$ - see front matter
PII: S0167-6393(02)00081-X

mit hoher Wortverständlichkeit synthetisiert werden kann und einen angenehmen Eindruck hinterlässt. Mit diesen ermutigenden Ergebnissen haben wir nunmehr ein anwenderfreundliches TTS-System mit Emotionen für die unmittelbaren Bedürfnisse Stimmloser entwickelt. Dieser Artikel beschreibt die vorgeschlagene Methode, die Konfiguration und die akustischen Eigenschaften der Corpora sowie die Ergebnisse der Erkennungsversuche.

Résumé

Nous proposons une nouvelle façon de produire synthétiquement la parole émotionnelle par un systéme de synthé se vocale par concatenation basé sur le corpus (ATR CHATR) utilisant le corpus de la parole émotionnelle. Avec notre méthode, la parole–avec l'emotion qui convient–peut être produit synthétiquement en changeant tout simplement entre des bases de données de source crées par le corpus. Nous avons evalué notre méthode en creant trois genres de corpus de la parole émotionnelle (la colére, la joie, et la tristesse) extraits des enregistrements d'un homme et d'une femme qui parlent Japonais. Les caractéristiques acoustiques de chaque corpus ne sont pas les mêmes et sont reconnaissables par émotion. Les caractéristiques acoustiques de chaque parole émotionnelle produit synthétiquement par notre méthode montrent des corrélations évidentes avec les caractéristiques acoustiques de chaque corpus. Des expériences perceptuelles utilisant la parole produit synthétiquement indiquent que notre méthode réussit à produire synthétiquement la parole émotionnelle de maniére reconnaissable. Nous avons evalué davantage l'intelligibilité et l'impression generale que notre méthode a fait sur les auditeurs. Les résultats montrent que la méthode proposée peut produire synthétiquement la parole avec un niveau élevé d'intelligibilité et donne une impression favorable. Avec ces résultats encourageants nous avons developpé un systéme TTS valable avec la capacité d'émotion pour répondre aux besoins immédiats des individus qui ne peuvent pas parler. Cet exposé décrit la méthode proposée, les caractéristiques acoustiques et de conception du corpus, et les résultats des évaluations perceptuelles.

Keywords: Emotion; Natural speech; Corpus; Source database; Concatenative speech synthesis

1. Introduction

Speech is the easiest way to convey intention, and it is one of the fundamental methods of conveying emotion, on a par with facial expression. Moreover, emotion plays an important role in communication (Guerrero et al., 1998). It is not difficult to imagine that if an individual loses both their ability to speak and their means of expressing their emotions, either vocally or even physically, due to paralysis, his or her life would be very isolated and depressing. Our research is motivated by the desire of patients suffering such hardships to expressively convey their intentions on their own. Several systems have reached the test stage, but no system is yet available on a commercial basis. Therefore, our goal is to develop a workable text-to-speech (TTS) communication-aid that can help express users' intentions and emotions.

Studies on emotional speech have verified that there are strong correlates between speech and emotion. Major findings in this area are described in key review articles of Davitz (1964a), Scherer (1986), Murray and Arnott (1993), and Cowie et al. (2001). The most commonly studied emotions in related studies are chosen from the 'basic emotions' such as happiness, sadness, fear, disgust, anger, and surprise (Murray and Arnott, 1993). Likewise, the acoustic parameters measured have been fundamental frequency (F_0), duration, energy, and, more recently, spectral characteristics and voice source parameters. Studies show that acoustic cues to emotion can be reliably found in natural speech and that listeners can also perceive emotion from speech.

Experiments using formant or LPC re-synthesis have shown that modifying acoustic parameters can produce recognizable emotional speech (e.g., for formant synthesizers, Carlson et al., 1992; and for LPC synthesizers, Ichikawa et al., 1967; Kitahara and Tohkura, 1992). Two groups, Murray and Arnott (1995) and Cahn (1989) have developed TTS synthesis systems for application to emotion. The system developed by the former is

named HAMLET, and the latter the Affect Editor. Both use DECtalk, a formant synthesizer. Recognition rates for perceptual tests with HAMLET indicated that emotions were recognized significantly for emotive texts. For the Affect Editor, all emotion types were recognized above chance level. HAMLET was later implemented in CHAT, a communication assistance system equipped with pre-composed phrases and a text-input acceleration function (Murray et al., 1991). Limitations to approaches using formant and LPC synthesizers include the facts that the acoustical representation of emotion still requires careful and heuristic adjustment and that a "non-human" voicing source is perceived as different from the human voice.

More recently, studies using speech waveform concatenation by uniform-sized units such as diphones have become popular. Signals are modified according to an emotion-dependent model using signal processing such as pitch synchronous overlap and add (e.g., Bunnell and Hoskins, 1998; Katae and Kimura, 2000; Murray et al., 2000). The results of these studies indicated that affective manipulations were possible but that the speech modification process introduced some distortion.

Using CHATR, a corpus-based concatenative speech synthesis system developed at ATR, we proposed a new approach to synthesizing emotional speech by creating emotional speech corpora. As a first trial, we have created corpora of three kinds of emotion: *anger*, *joy*, and *sadness* of a male and a female speaker of Japanese. The result of an acoustic analysis and a perceptual evaluation showed that each corpus was acoustically different from the others and emotionally identifiable, as was the speech synthesized by our method. Further experiments showed that our method could synthesize speech that provided high intelligibility and gave a favorable impression to listeners. Based on these results, we developed a workable TTS system with emotion for nonspeaking individuals. This paper provides a detailed description of that work.

In this paper, *emotion* is used in its broad definition, which involves the abrupt change to a stable state of mind more often referred to as "mood". Also, *emotional* is used interchangeably with *emotive* and *affective* to mean "being capable of expressing".

2. CHATR: the TTS synthesis system used in our approach

CHATR is a natural-speech re-sequencing synthesis system that uses a large library of source units (hereafter, a source database) retaining natural variation for unit selection to reproduce speech with the original phonetic and prosodic characteristics of the speaker (Campbell, 1997a,b; Campbell and Black, 1997). It works on UNIX, Linux and MS Windows (CHATR98).

2.1. Source database creation in CHATR

Creating a source database for CHATR is an off-line one-time process. The readings of text materials from one speaker (a speech corpus) are digitally recorded and stored in a PC. Then disfluencies and redundancies within the speech are eliminated, and the speech waveforms are transcribed. The speech database is stored externally to the synthesizer. In addition, an inventory file for access to the database is created according to the following procedure: (1) converting an orthographic transcription of the corpus texts to an equivalent phonemic representation, (2) aligning the phonemes to the waveform to provide a start time for each phone-based unit so that prosodic feature can be measured, and (3) producing feature vectors for each unit. Features included are phoneme label, starting time, duration, F_0, probability of voicing and RMS energy. Phoneme labels of neighboring units are easily identified in the inventory from the starting time information. An example of an inventory file is given in Fig. 1. CHATR then calculates the weight vectors of each feature per unit to ensure optimal candidate in the unit selection process.

The ideal size and balance of a corpus to serve as a source for the speech database has not yet been determined, but the ATR phonemically balanced text corpus (hereafter ATR 525 sentences), a corpus with some supplements to the original ATR 503-sentence corpus developed by Abe et al.

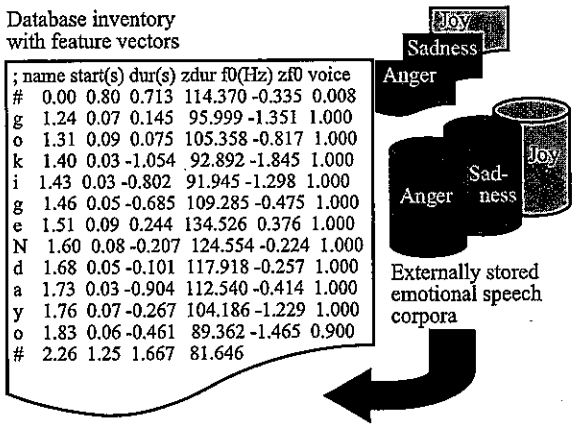


Fig. 1. Database inventory creation. start(s): starting time, dur: duration, zdur: z-score of duration, f0: fundamental frequency (F_0), zf0: z-score of F_0, voice: probability of voicing.

(1990), has proved to be sufficient, and is used here as the baseline reading material. This corpus is a list of isolated sentences. The profile of this corpus is 525 sentences, 16,612 moras, 31,053 phonemes, and 403 biphone variations (abbreviated as "525S" in Table 2). Approximate reading time for this corpus is about an hour.

2.2. CHATR TTS synthesis process

After texts are typed in, CHATR takes three steps in the real-time online TTS synthesis process: (1) text analysis (text-to-phoneme conversion, accent tagging, break indexing), (2) prosodic prediction, and (3) unit selection. Fig. 2 shows the process flow of CHATR.

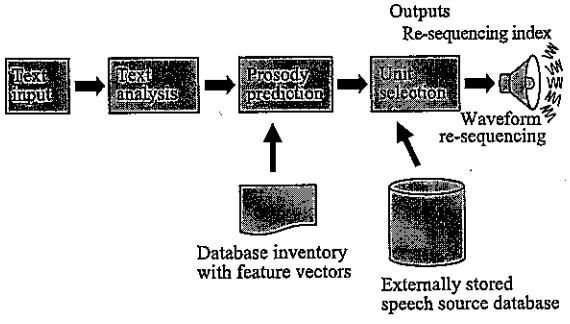


Fig. 2. CHATR process sequence.

2.3. Prosody prediction in CHATR

In the standard CHATR, F_0 and duration are predicted. The CHATR synthesis process makes use of linear regression (LR) models trained with a reading of the ATR original corpus and some other materials, produced by a professional male announcer of standard Japanese in his normal reading manner. As for F_0, three LR models for each mora are trained. These predict F_0 values at the start, mid-point and end of each mora from Tones and Break Index (JtoBI) labels (Campbell, 1996; Black and Hunt, 1996). (A mora is a unit of pronunciation in Japanese corresponding to a single vowel or a consonant-vowel combination.) LR models for duration predict values for each unit in terms of normalized z-scores. These are converted to absolute durations in milliseconds by reference to a table of means and standard deviations (SD) stored as a part of the synthesis speaker database. In this way, standard patterns of phonemic timing are mapped onto speaker-specific characteristics of speaking rate.

CHATR then maps the predicted values to the source database by means of z-score normalized values using the mean and SD of F_0 and duration per segment as observed for that source database (Campbell and Black, 1997). The predicted prosodic patterns can be mapped to any source database by this transform, enabling the production of speaker-independent prosodic phrasing. The advantage of this method is that the predicted pattern serves as a guideline representing the shape of the intonation patterns, and thus by selecting units closest to that pattern, CHATR finds the optimal sequence of speech units for that speaker from the available units in the database.

2.4. Unit selection algorithm of CHATR

In the unit selection process, CHATR first lists up candidate units for concatenation from the database inventory by looking up the feature vectors and the neighboring units' information. Then, in order to create a new speech waveform representing the target utterance, it selects the optimum unit and thereby to produce an index listing the optimum sequence of units to be randomly accessed

from the externally-stored natural speech corpus. The unit selection is made according to (a) a target cost, the degree to which they fit the target specification, and (b) a concatenation cost, the degree to which they can be imperceptibly concatenated.

3. A new approach to synthesizing emotional speech

We proposed an approach to synthesizing emotional speech by creating an emotional speech database (Iida et al., 1998). After evaluating its validity, this method was implemented in a TTS system with emotion using a graphical user interface (GUI) to assist users with the choice of speaker and/or emotion selection and other TTS commands (Iida et al., 2000). A description of its GUI is given in Section 7.2.

3.1. Basic concept of the proposed approach

The prosody prediction and unit selection algorithm of CHATR is designed to make the best use of the natural variation in the source database. Therefore, the speech synthesized by CHATR reflects the nature of the original database. Taking this as a starting point, we assumed that if we could create an emotional speech corpus with natural phonetic and prosodic characteristics appropriate to that particular emotion, then we could synthesize speech reflecting the characteristics of that emotion by synthesizing from that corpus using CHATR. Furthermore, if the above assumption were valid, we could synthesize speech with varied emotion as long as we could make an emotionally identifiable and distinguishable corpus.

When several emotions are to be synthesized in one production, as when a speaker changes his/her speaking-style during an utterance, speech with the appropriate emotion is synthesized by simply switching from one source database to another according to the user's commands, which are internally embedded in the text (cf. Section 7.3).

3.2. Process flow of emotional speech synthesis

Fig. 3 illustrates the process flow of the system with three speech databases for anger, joy, and sadness, as examples. The upper layer illustrates the standard CHATR TTS synthesis procedure and the lower layer shows the emotion database selection with source databases for emotional speech.

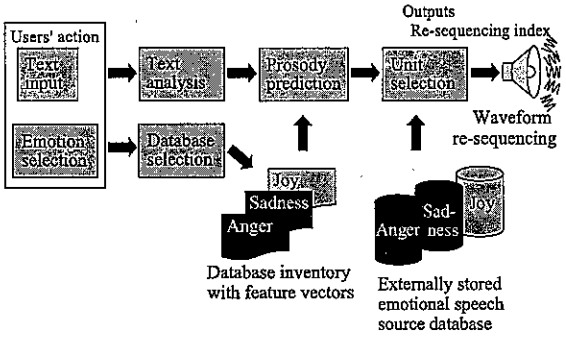


Fig. 3. Process flow of proposed method.

3.3. Prosody prediction for emotional speech

In this study, no emotional-dependent prosody prediction model was used. Developing emotion-dependent models is planned for future work, but here, we used the standard prediction model currently available in CHATR. As described in Section 2.3, this model is trained by speech read with a normal reading manner that can be interpreted as read in an emotionally neutral tone. As mentioned, a z-score transform (the mean and SD of F_0 and duration of the source database are encoded as z-scores) is used to determine the prosodic contours in CHATR. In other words, CHATR predicts the prosodic patterns and leaves the range to be determined by the acoustic characteristics of the source database. Thus, F_0s and durations of the synthesized speech deviate from predicted trajectories of the normal reading manner of the model speaker according to F_0 and duration variations of the source database. In this way, we hypothesized that the prosody of emotional speech can be, to a certain extent, implicitly decided by the acoustic characteristics of the source database without any further signal processing. For example, the same F_0 pattern should become more compressed with lower values when units are selected from a database that has a narrower F_0 range (e.g. in speech

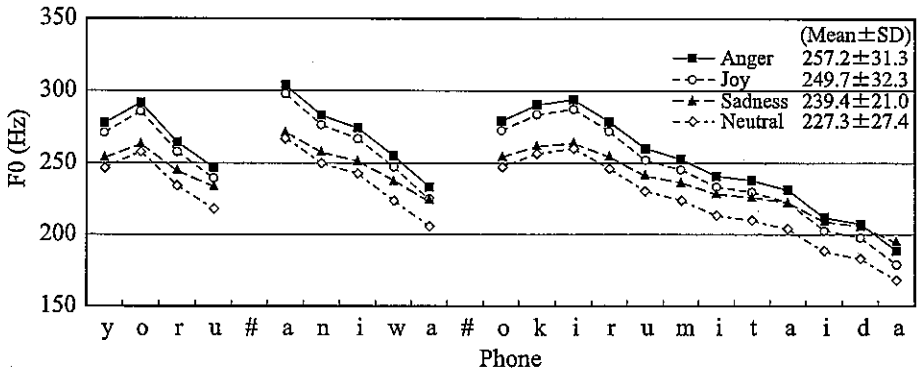


Fig. 4. Target F_0 per phone-based unit for "Yoru, aniwa okiru mitaida (My brother seems to wake up at night)" using for kinds of corpus: anger, joy, sadness, and neutral of the female speaker.

expressing sadness). This hypothesis has been confirmed in this study by observing speech synthesized by the proposed method. Target F_0 trajectories for four different speaking styles are shown in Fig. 4. The corpora used for this experiment were anger, joy, and sadness of the female speaker created in this research, and as a reference, a speech corpus of a reading of the ATR 525 sentences by the same speaker (the *neutral corpus*, here after). The target durations of the four different kinds of speech can be seen in Fig. 11(b) in Section 6.1.1. The description of how emotional corpora were made is given in Section 4, their acoustic profiles in Table 3 of Section 5.2, and the description of the neutral corpus is given in Section 6.1.1. This corpus was recorded with an interval of four years after the recording of emotional speech corpora, to meet the need of users to add neutral speech (cf. Section 6.2.4). The sentence used for synthesis is "Yoru, aniwa okiru mitaida (My brother seems to wake up at night)". The following can be observed from Figs. 4 and 11(b): (1) the shapes of the target F_0 trajectories of the four speech samples are similar, (2) F_0 trajectories for anger and joy are plotted in higher frequency regions while those of sadness and neutral speech are plotted in lower frequency regions, (3) the duration per unit of anger and joy are longer than those of sad speech and neutral speech. The mean and SD of F_0 per unit for each speech sample were angry: 257.2 ± 31.2 Hz, joy: 249.7 ± 32.3 Hz, sad: 239.4 ± 21.0 Hz, and neutral: 227.3 ± 27.4 Hz. The mean and SD of duration were angry: 97.3 ± 57.4 ms, joy: 90.7 ± 59.3 ms, sad: 83.1 ± 60.1 ms, and neutral: 84.4 ± 59.7 ms. Further description with a comparison between target and selected units is made in Section 6.1.1.

4. Creating corpora of emotional speech

In the proposed system, the synthesis of good emotional speech is, to a large extent, underpinned by the quality of the corpus being used. To verify our assumption, we created three read-speech corpora from a male and a female speaker of Japanese under three different emotional states (anger, joy, and sadness).

4.1. Emotional variations selected for this research

In order to determine what kinds of emotional expressions the target users need, we referred to their autobiographies (Todoroki, 1993; Ohira, 1995; Kamimura, 1990) and then interviewed them. The following emotions emerged: anger, rage, disgust, unwillingness, gratitude, happiness, pleasantness, elation, sadness, disconsolation, loneliness, and anxiety.

In the proposed method, it would be possible to create a corpus for each emotion listed above if there is enough speech with the appropriate emotion. However, a large volume of data is required to meet the size requirement for CHATR synthe-

sis, and it was very difficult to collect the necessary amount of data for each corpus. Therefore, we grouped the emotions as below according to similarity in the quality dimension (i.e., belonging to the same emotion family) by referring to the prototype approach (Shaver et al., 1987) as a first step.

Group 1. Anger, rage, disgust, unwillingness;
Group 2. Joy, gratitude, happiness, pleasantness, elation;
Group 3. Sadness, disconsolation, loneliness and anxiety.

The first group was labeled the *anger corpus*, the second the *joy corpus*, and the third the *sadness corpus* for the sake of convenience, but these are not necessarily equivalent to the same terms referring to the discrete emotions. Each corpus contains subordinate emotions as listed above similar in quality but different in intensity, and a wide diversity in acoustic characteristics can be found in each corpus. Fig. 5 is a two-dimensional diagram similar to Russell's circumplex model (Russell, 1989), where the dimension of each corpus is shown; joy is placed in the pleasure hemisphere, covering about 3/4 hemisphere from the highest arousal point, while anger is in the displeasure/high arousal quadrant and sadness is in the displeasure/low arousal quadrant.

4.2. Reading materials for the corpora

In order to synthesize speech with CHATR, we need a large volume of phonetically and prosodically well-balanced emotionally-colored speech data recorded in a studio-quality environment. We set the ATR 525 sentences with 31,053 phonemes (cf. Section 2.1) as a minimum baseline for collecting text materials to be read.

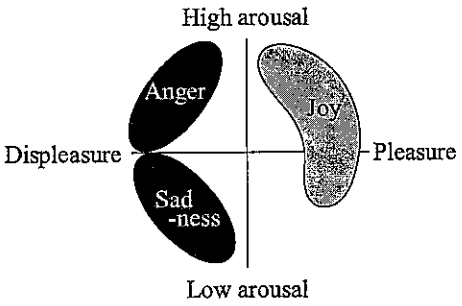


Fig. 5. Dimensions of each emotional corpus.

For this research, we designed a novel approach using monologues. The first author composed and collected monologues illustrating appropriate emotions for each emotional corpus from newspapers, web pages, and self-published autobiographies of disabled people. Only monologues that caused her to experience the target emotion were included in the text corpus of that emotion. Each is not just a list of sentences but a complete essay that has a story to tell. In this way, we assumed that speakers could more easily maintain a particular emotional state for a sustainable period of time. As a result, the three corpora contain completely different texts. Each monologue consists of about 37 sentences. An example sentence for each text corpus is shown in Table 1.

4.3. Speakers

In related studies, actors and announcers have been asked to simulate various emotions. However, depending on the differences in competence between actors, overacting or exaggeration of the actors may result in unnatural expression, which differs from speech produced by a speaker when experiencing a genuine emotion (Scherer et al., 1991).

In our research, non-professionals who have a good reputation for being vocally expressive were selected for both female and male speakers in order to avoid exaggerated expressions. A female graduate student with a standard Tokyo dialect (the first author) was selected as a speaker for the first trial and a male graduate student also from Tokyo was selected as a second speaker.

4.4. Emotion elicitation

Elicitation techniques such as the scenario approach (e.g., Scherer et al., 1991; Nagae, 1998), adding a leading paragraph (e.g., Fairbanks and Pronovost, 1939; Davitz, 1964b), and the dialogue approach (e.g., Ito, 1986) have been reported, but these were designed for experiments in which

Table 1
Example sentences from each text corpus

Emotion	Example texts	English translation
Joy	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 yatt-emirukarana! Iya, gokigendayo, kore!" (Adding new sentences expressing elation.)	Even a person like me with a body completely paralyzed can do it! "It is great! Just great! Hey, I can do things like this, too! Look at me, are you ready? See, I will do it again. Oh, it's absolutely fantastic!"
Anger	Dakara Yamamoto-san (Addressing), nando mo itteiru youni, anatawa honnin no manzoku no tameniha "keishiki wa nandemo ii" tte rikisetsu suru keredo, "houritsu ni mamotte morau" tameniha "keishiki wa nandemo ii" toiu wakeniha ikanai janaidesuka! (Emphasis)	Ok, Ms. Yamamoto, as I have been telling you a number of times, you are insisting that formality is worth nothing, but if you want the law to protect you, you just cannot say so. You need to register your marriage properly!
Sadness	Jibun de jibun no sei no jikan ga kimeraretara, donnani iidarou. Shi wo Kangaenagara ikasareru kotono tsurasa. Watashi no tameni kenmei ni akarukushite kureru shujin ya kodomo tachi no mae de nakukotomo dekizu, ii hahaoya wo enjiteiru. Demo hitori ni naruto setsunakute namida ga tomaranai. (Replacing a sentence contains the word "angry"—Sonna jibun ga haradatashii.)	How nice it could be if I would decide how long I lived. How hard it is to be alive thinking about death. I cannot cry in front of my husband and children who are trying hard to be cheerful. I act as a good mother but I get sad and cannot stop crying when I am left alone. (Replacing a sentence which means "I am mad at myself".)

Examples of text modification are underlined and modification tactics are shown in parenthesis.

only a small number of utterances are investigated. Therefore, a different approach, by way of eliciting by text materials, was carried out in this research.

4.4.1. Text modification

It is preferred that all utterances from a speaker be spoken under the target emotional state. However, in the monologues there are some descriptive sentences explaining the background information and situation, which are presumably uttered in a neutral or near-neutral tone of voice. Therefore, texts were modified with the authors' permission. Modifications to text materials were made in order to shift the base emotion to a low-arousal state with respect to the target emotion and also to stimulate the speaker to produce some high-arousal representation of the target emotion. Tactics of text modification are listed below, and examples of text modifications are underlined in Table 1.

1. Adding new sentences to build up a high arousal of the target emotion.
2. Emphasizing the target emotion by adding emotive modifiers.
3. Keeping the rhythm and tempo to enhance the speaker's emotive state by addressing the interlocutor, through insertion of phrases typical to each emotion in appropriate places.
4. Replacing expressions that might stimulate other distinctive emotions with expressions more likely to maintain the target emotion without changing the meaning drastically.
5. Highlighting the high arousal expressions in the anger and joy corpora by use of exclamation marks (no symbol of exclamation is used in traditional Japanese orthography).

4.4.2. Elicitation during recording

All corpora were recorded in a sound-treated studio at ATR. Each emotional corpus was recorded in one session, and the speaker took a break between sessions. Each session lasted for about 90 min, including a brief rest and talk with the experimenter located outside the sound-treated room. No special elicitation was attempted for the female speaker since she was the one who collected, composed and modified the monologues. Consequently, the text materials were familiar to her before the recording.

As for the male speaker, the experimenter (the first author) attempted to elicit the target emotion from a male colleague during the recording. A brief introduction to the authors and background

information on the texts were presented at the beginning of each recording session. Occasionally during the recording, the experimenter started up a conversation with topics intended to help the speaker experience the target emotion. Typical topics of these conversations were as follows. For joy, a fun experience in common between the speaker and the experimenter, for anger, a police officer tagging him for speeding, and for sadness, his recent bereavement. The total time spent in the elicitation conversation between the speaker and experimenter was approximately 30 min, and the speaker's free talk was about 15 min for each corpus. A recording of the conversation was not included in the corpora to protect his privacy.

4.5. Size of each speech corpus

All texts were read and digitally recorded on Digital Audio Tape. They then were downsampled and stored in a PC as 16-bit Microsoft wav format files at 16-kHz sampling. A total of six corpora (i.e., anger, joy, and sadness for both female and male) were created. All corpora except for the male sadness corpus were larger than the ATR 525 sentences (cf. 2.1) in terms of the total number of phonemes. The sizes and other profiles of each emotional corpus are given in Table 2 with the profile of the ATR 525 sentences for reference. The difference in data size between the sadness corpus of the male and female speaker is due to the difference in the amount of reading.

The phonemic balance of each corpus was neither controlled nor adjusted in the design stage, giving priority to maintaining the naturalness of the reading materials. The total number of biphone variations of emotional corpora varied from 345 for the female sad corpus to 402 for the male joy corpus. Compared to 403 for the ATR 525 sentences, the female sadness corpus is about 86%. The prosodic variation was not controlled or adjusted in the read materials since this methodology itself is not yet established.

5. Characteristics of the emotional speech corpora

An important prerequisite to this research is that the created speech corpora sound emotional and different from each other. Both aspects were examined by the subjects' evaluations of text materials and speech materials. Then an acoustic analysis was performed to confirm the latter.

5.1. Evaluation of emotional speech corpora

One popular method to evaluate whether subjects could correctly recognize the intended emotion of speech stimuli is a forced choice test. Commonly, distracters, or items such as "other", are included in the selection list (e.g., Fairbanks and Pronovost, 1939; Murray and Arnott, 1995). In our evaluation, we performed the test among joy, anger, and sadness and did not include any distracters, since the aim of the test was to see if the subjects could differentiate the three kinds of emotional speech. We assumed that "distinguishable among the choices" is the first step and a sufficient requirement for a system operated under a consensus between the user and

Table 2
Size of source database created from emotional corpus (F stands for Female, M for Male)

	Monologues	Sentences	Moras	Phonemes	Biphones	File size (MB)	Length (min:s)
Joy (F)	12	461	21,680	40,928	377	110.0	51:12:00
Anger (F)	15	495	21,005	39,171	349	98.3	57:18:00
Sadness (F)	10	426	16,612	31,840	345	93.3	48:36:00
Joy (M)	12	461	21,709	39,878	402	84.1	43:48:00
Anger (M)	15	495	21,231	38,360	388	83.6	43:33:00
Sadness (M)	9	343	14,655	27,302	383	49.8	31:09:00
525S	–	525	16,612	31,053	403	81.0	42:11:00

525S is shown as a reference.
Figures for Biphones are the total numbers of variations and figures for the rest are accumulating totals.

interlocutor over the number of available emotions.

5.1.1. Evaluation of text material

Eighty-seven university student volunteers (63 males, 24 females) were asked to read and judge the emotional type of each monologue from all corpora (37 monologues in total) from a forced choice selection of anger, joy, and sadness. Each essay was read by at least two students. As a result, all texts but two were correctly judged as representing the emotion types that the experimenter had classified for them.

5.1.2. Perceptual evaluation of recorded speech

To see whether the speech of the created corpora expresses the intended emotion, a forced choice test among anger, joy, and sadness was again performed. Tests were performed separately since the date of completion of the speech corpora differs, with the female speech corpora created first, followed by the male speech corpora. For the emotional speech corpora of the female speaker, an additional question asked whether the emotion type of the speech stimuli could be judged from the textual content alone.

Subjects were 29 university students (15 males, 14 females) for both experiments. In order to minimize the contextual interference, the entire waveform in each corpus was segmented into individual sentences with each representing one speech stimulus. While semantic textual information is naturally retained in stimuli thus segmented at the sentence-level, it was decided not to use smaller segmentation units in order to retain sentence-level prosodic characteristics. Fifty speech stimuli were randomly selected and presented twice to two subjects via a notebook PC (Toshiba Dynabook SS3300/CPU266 MHz, MS Windows 98). All stimuli were stored in a PC as 16-bit Microsoft wav format files at 16-kHz sampling. Subjects heard the speech through headsets (SONY MDR-NC20). The same PC was used to collect all responses from the subjects.

As shown in Figs. 6 and 7, sad speech was identified most successfully among the three kinds of emotional speech of both speakers. The results for the female speech were anger: 86%, joy: 80%, and sadness: 93%; results for the male speech were anger: 77%, joy: 70%, and sadness: 95%. We performed a test of difference between means for a binomial distribution by setting a null hypothesis of identification rate for each emotion at a chance level of 33%. The result rejected this hypothesis at a significance level of $p < 0.01$. As for the additional question asked for the female speaker corpora, 47% of the total responses were marked to indicate that the subjects' judgments were made from the textual content. The result of the identification test showed that the intended emotion was recognized at an incidence significantly greater than chance, but we cannot exactly know how much was recognized solely from the speech, due to the nature of the corpora.

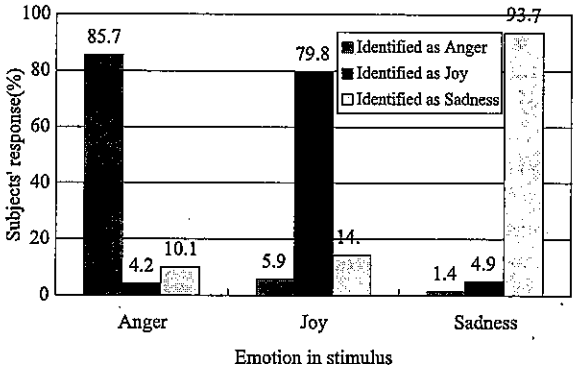


Fig. 6. Types of emotion identified for speech stimuli from emotional speech corpora of female speaker.

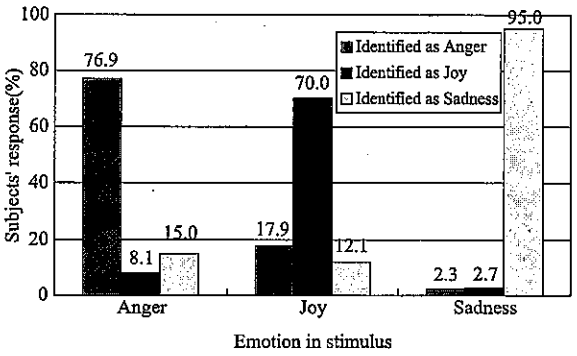


Fig. 7. Types of emotion identified for speech stimuli from emotional speech corpora of male speaker.

5.2. Acoustic characteristics of speech corpora

Using the emotional corpora of both speakers stored in a PC, prosodic parameters such as F_0, duration, and RMS energy were measured. In addition, the formants of the steady-state vowels selected from the female speaker's data were measured. Since the three types of emotional corpora contain completely different texts and no reference texts were included, the standard approach of comparing parameters for an identical phone sequence within identical sentences could not be used. Instead, a statistical approach was used by taking advantage of the large amount of data to produce a global profile of each emotional corpus. The results confirmed the finding of earlier studies that F_0 and duration were important parameters for differentiating emotion (e.g., Scherer, 1986; Murray and Arnott, 1993; Banse and Scherer, 1996; Mozziconacci, 1998).

5.2.1. Fundamental frequency

F_0 was measured across the entire waveforms of each corpus using ESPS's get_f0 command (Entropic, 1996), which implements an F_0 estimation algorithm using the normalized cross correlation function with a frame step of 10 ms and a correlation window of 7.5 ms. We used only those units with positive voicing values for analysis.

Mean and SD of F_0 are shown in Table 3 along with those of duration and RMS. The mean and SD (written as mean $\pm$ SD) of F_0 for each female speech corpus were anger: 255.8 ± 52.1 Hz, joy: 249.1 ± 49.3 Hz, and sadness: 235.7 ± 34.5 Hz, and that for each male speech corpus was anger: 174.8 ± 38.9 Hz, joy: 161.1 ± 36.1 Hz, and sadness: 124.9 ± 20.8 Hz. For both speakers, mean F_0 of the sadness corpus was the lowest and its SD was the smallest. This result supports findings in earlier studies. Furthermore, the sadness corpus of the male speaker has substantially lower mean F_0 than those of his anger and joy corpora, whose values are close to one another. Means comparisons by ANOVA showed that the F_0 of the three emotions were significantly different from one another at $p < 0.05$ for both male and female speakers. The histograms in Figs. 8 and 9 both show the narrowest range for the sadness corpus and the broadest for the anger corpus. The distributions of all corpora do not have a Gaussian shape.

5.2.2. Segmental duration

For the duration analysis, we measured the duration of vowels and voiced consonants. Their location in the speech waveform was identified by time index and phoneme labels that were created by auto-labeling and a succeeding manual check when creating the CHATR source database.

The mean and SD of the unit durations for each of the female speech corpora were anger: 67.3 ± 29.5 ms, joy: 65.3 ± 30.1 ms, and sadness: 74.6 ± 32.8 ms, and the corresponding values for the male speech corpus were anger: 60.7 ± 26.8 ms, joy: 58.6 ± 25.3 ms, and sadness: 56.2 ± 20.8 ms. ANOVA showed that the means of the three emotions were significantly different from one another at $p < 0.05$ for both speakers. Earlier studies reported that the segmental duration was longer in sad speech. This tendency can also be observed in the data of the female speaker, where mean duration of the sadness corpus is substantially longer than those of her anger and joy corpora, which

Table 3
Prosodic parameters of emotional speech corpora

Corpus		F_0 (Hz)	Duration (ms)	RMS energy (dB)	RMS for selected vowel (mean $\pm$ SD) (dB)
Female	Anger	255.8 ± 52.1	67.3 ± 29.5	60.5 ± 4.9	68.8 ± 3.6
	Joy	249.1 ± 49.3	65.3 ± 30.1	61.2 ± 4.7	68.2 ± 3.4
	Sadness	235.7 ± 34.5	74.6 ± 32.8	61.1 ± 4.6	66.9 ± 3.4
Male	Anger	174.8 ± 38.9	60.7 ± 26.8	57.5 ± 7.7	
	Joy	161.1 ± 36.1	58.6 ± 25.3	57.9 ± 7.5	
	Sadness	124.9 ± 20.8	56.2 ± 20.8	60.0 ± 6.3	

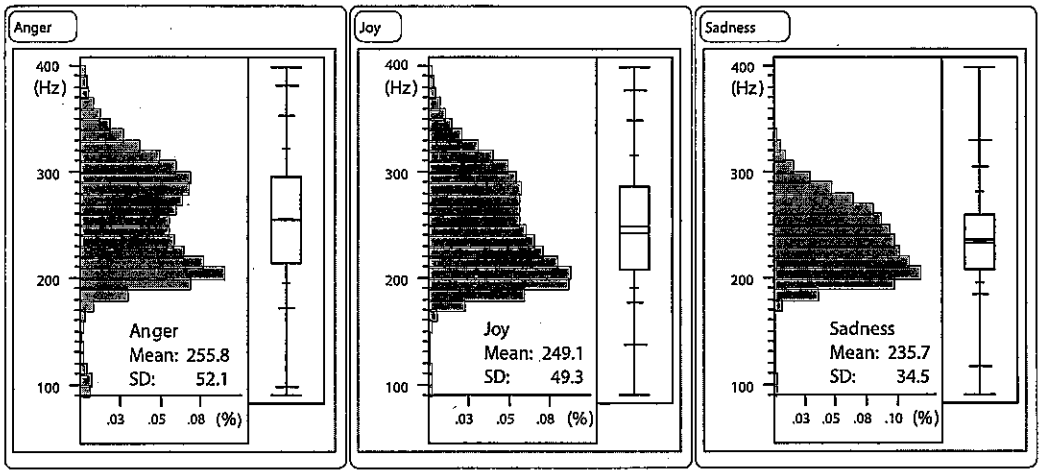


Fig. 8. F_0 of female speech corpora (from left to right: anger, joy and sadness). Histograms showing the proportion of total segments (X) by frequency (Y). The middle of the quantile box is the median, and the 25th and 75th quantiles are the ends.

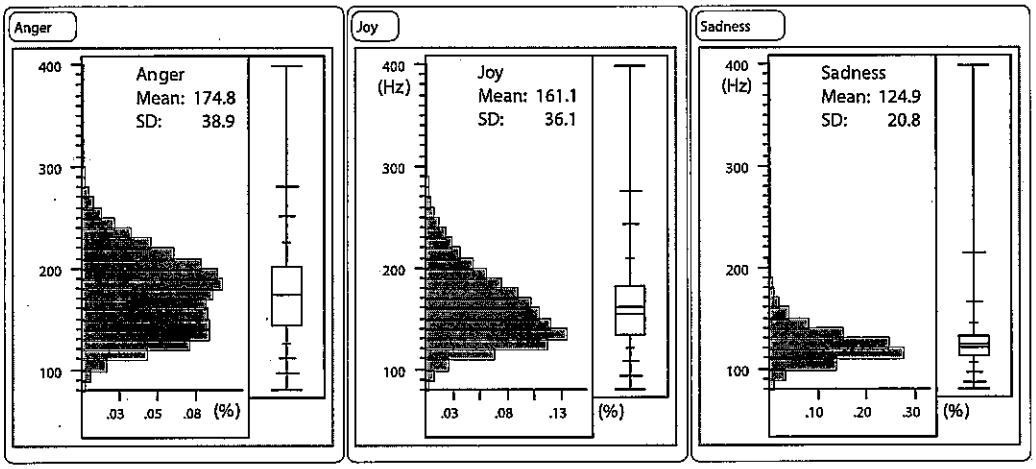


Fig. 9. F_0 of male speech corpora: histograms for all voiced segments.

have values that are also much closer together. However, this tendency was not observed in the male speaker's data, where the mean duration is the shortest for his sadness corpus. In addition to the mean values just discussed, the SD of the sadness corpus of the male speaker is much smaller than his other two corpora while the SD of the sadness corpus of the female speaker is the largest. These SD values imply that the durational variation among units is smaller in the male speaker's sad speech and larger in the female speaker's sad speech.

5.2.3. Root mean square energy

The root mean square (RMS) energy of the entire waveform of each corpus was measured by get_f0, and the logarithm of RMS ($20 \times \log_{10}$ (RMS), in dB) of each frame was computed. In get_f0, the RMS value of each record is computed based on a 30 ms hamming window with its left edge placed 5 ms before the beginning of the frame.

The mean and SD of energy for each female emotional corpus were anger: 60.5 ± 4.9 dB, joy: 61.2 ± 4.7 dB, and sadness: 61.1 ± 4.6 dB, and that

for each male speech corpus was anger: 57.5 ± 7.7 dB, joy: 57.9 ± 7.5 dB, and sadness: 60.0 ± 6.3 dB. Unexpectedly, the RMS energy for the sad speech corpus was not the weakest of the three corpora for both speakers. Here, we must keep in mind that what we measured was not the actual speech but a recording of the speech. We had to set the recording volume to a higher level during the recording of the sadness corpus than during the recording of the anger and joy corpora, since currently in our method, no drastic change is possible in the intensity level when the speech is synthesized. This means that what was measured does not reflect the absolute value. However, it is assumed that the relative values within each corpus are preserved and that speech with variation close to natural can be produced when the units are optimally selected.

As a reference closer to the actual RMS representation, we measured the RMS energies of the steady-state vowels selected from the female speech corpora. Fifteen examples of each of the five vowels /a, e, i, o, u/ were selected from each of three emotional corpora on the basis of longest-duration vocalic nuclei that also carry perceptually identifiable characteristics of the intended emotion. Thus a total of 15 nuclei × 5 vowels × 3 emotions = 225 vowel segments were used for analysis. The mean and SD of energy were anger: 68.8 ± 3.6 dB, joy: 68.2 ± 3.4 dB, and sadness: 66.9 ± 3.4 dB. ANOVA results showed that only the means of the anger–sadness pairs were significantly different from each other at $p < 0.05$.

So far, we could not find evidence that energy profiles differ according to emotion from the global energy profiles (means and SDs of each corpus as a whole) analyzed in this study. The reason might well be attributed to the way energy levels are measured as described above. More careful measurement, for example, the direct measurement of energy levels of speech using a sound level meter, is included in our future work. Some studies have reported that effects of emotions appeared at local energy profiles (e.g., Kitahara and Tohkura, 1992; Takeda et al., 2000). It would certainly be interesting to focus specifically on energy levels of sub-segments of phones such as the onset burst of plosives and high frequency components of fricatives of accented syllables of focused words.

5.2.4. Formants

The vowel formants of the female speaker were analyzed from the same set of vowel segments used in the RMS energy analysis (cf. Section 5.2.3). These formants were measured by Mokhtari et al. and formed the basis of our study of articulatory correlates of emotion variability (Mokhtari et al., 2001). A semi-supervised, cepstrum-based method was used to identify the five steadiest states, i.e., consecutive frames in each vowel nucleus, and the first four formants were then selected from among the poles of a selective linear-prediction analysis.

Fig. 10 shows the resulting F_1–F_2 vowel space of the female speaker. The relative location of each vowel within that space agrees well with the general F_1–F_2 distribution of the five Japanese vowels (e.g., Keating and Huffman, 1984), where a high vowel /i/ is more lax than English /i/, and /u/ is less rounded than English /u/. In F_1–F_2 spaces, both are located in more centralized positions. The two-sigma ellipses indicate the size and direction of the variability across all measured data for each

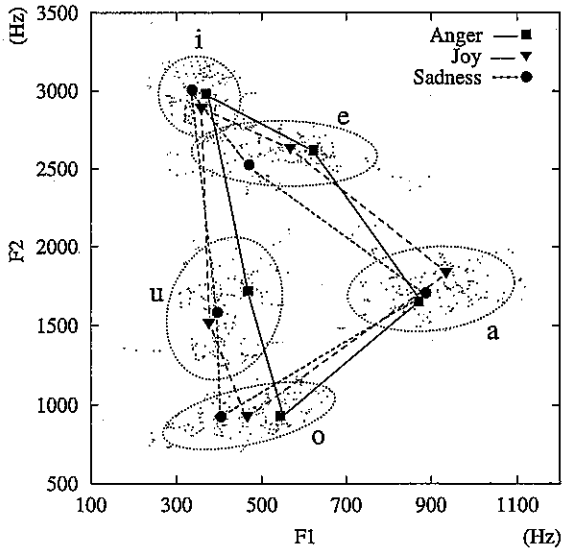


Fig. 10. F_1–F_2 diagram for five vowels of female speech corpora. Two-sigma ellipses indicate the spread over all data per vowel.

vowel. As can be seen, /i/ has the tightest distribution, /e/ and /o/ are relatively consistent in F_2, and /u/ and /a/ have the widest range of values.

The formant variability is investigated from two viewpoints based on Makino et al. (1989). The first is that the formant values of all vowels tend to shift to higher frequencies with higher F_0, especially across speaker-groups. The second viewpoint is that faster speaking-rates tend to yield reduced vowels having more centralized distributions in the F_1–F_2 plane. Table 4 shows the mean values of F_0, F_1, F_2, F_3 and vocalic segment duration for each of the five vowels and three emotions separately. The mean F_0 decreases in the order of anger > joy > sadness, while duration has the reverse trend. To confirm both viewpoints, we compared F_1 and F_2 values for each pair of anger–sadness, anger–joy and joy–sadness. A circle in Table 5 indicates that the vowel with higher F_0 has the higher F_1 and/or F_2. Likewise, a circle in Table 6 indicates that the vowel spoken with faster speaking rates has more centralized F_1 and/or F_2. The factors supporting both viewpoints were best observed when comparing anger and sadness, where larger differences in both F_0 and duration exist.

Table 4
Mean F_0, F_1, F_2, F_3 and duration of vowels in the female speech corpora

Vowel	Emotion	F_0 (Hz)	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)	Duration (ms)
/a/	Anger	260.4	871.2	1654.5	3135.2	84.2
	Joy	255.6	938.4	1842.3	3209.6	90.1
	Sadness	243.7	887.5	1702.0	3095.6	97.5
/e/	Anger	264.9	622.9	2619.7	3265.6	79.3
	Joy	254.6	567.5	2642.3	3308.5	81.5
	Sadness	241.2	472.8	2533.3	3228.2	89.4
/i/	Anger	257.2	369.2	2981.1	3565.7	67.1
	Joy	248.8	356.6	2901.9	3510.7	67.7
	Sadness	238.6	336.4	3004.8	3573.5	72.4
/o/	Anger	259.4	544.6	931.9	3108.8	75.0
	Joy	254.2	462.6	933.2	3136.0	77.3
	Sadness	239.9	405.5	929.6	3210.9	84.6
/u/	Anger	262.1	467.9	1719.5	3091.1	59.3
	Joy	259.9	376.8	1522.4	3067.9	60.5
	Sadness	243.7	393.9	1585.6	3070.7	64.6

Table 5
Data table to confirm viewpoint 1 (see text)

	Vowel	Anger–Sadness	Anger–Joy	Joy–Sadness
F_1	/a/	×	×	○
	/e/	○	○	○
	/i/	○	○	○
	/o/	○	○	○
	/u/	○	○	×
F_2	/a/	×	×	○
	/e/	○	×	○
	/i/	×	○	×
	/o/	○	×	○
	/u/	○	○	×

A circle indicates that the value for the corpus written on the left side of each pair is higher in frequency.

Table 6
Data table to confirm viewpoint 2 (see text)

	Vowel	Anger–Sadness	Anger–Joy	Joy–Sadness
F_1	/a/	○	○	×
	/e/	×	×	×
	/i/	○	○	○
	/o/	○	○	○
	/u/	○	○	×
F_2	/a/	×	×	○
	/e/	×	○	×
	/i/	○	×	○
	/o/	○	×	○
	/u/	○	○	×

A circle indicates the following: For F_1, the value on the left is higher for /i/, /o/, /u/, and lower for /a/, /e/. For F_2, the value on the left is higher for /a/, /o/, /u/, and lower for /e/, /i/ (these are the characteristics of the centralization).

While the implications of these findings must be investigated in greater detail, they can be briefly interpreted as follows: emotional speech (or speaking style) spoken at higher pitch and faster speed (such as in speech expressing anger) has a narrower (or more reduced) F_1–F_2 vowel space, which is also shifted to higher frequencies.

6. Evaluation of the synthesized speech

Using each emotional source database in turn, synthesized speech was produced by our method.

An analysis of selected units was conducted and perceptual evaluations were performed on the synthesized speech. The common conditions for the perceptual experiments are as follows. All subjects were unpaid volunteers with no hearing disorders, and all speech stimuli were saved as 16-kHz, 16-bit wav format files. Each speech stimulus was selected randomly and presented twice to subjects via a notebook PC (Toshiba Dynabook SS3300/CPU266 MHz, MS Windows 98). The amplifier settings were set to a comfortable level for all subjects and for all speech stimuli. Subjects heard the speech through headsets (SONY MDR-NC20) unless specified. Subjects were asked either to choose from the selection list or write their responses on answer sheets.

6.1. Determining whether synthesized speech reflects speech corpora characteristics

The prosodic parameters of the synthesized speech were analyzed, and an emotion identification test was performed. The objective was to evaluate whether our method could reproduce the characteristics of the original emotional speech corpus.

6.1.1. Prosodic parameters of synthesized speech

In order to examine how selected units differ from their targets, we, first, measured global acoustic profiles of speech samples synthesized by our method. We analyzed the means and SDs of F_0 and duration of thirty synthesized speech samples (5 sentences × 3 emotions × 2 speakers) used in the emotion identification test described in Section 6.1.3. The sentences used are listed in Table 8 in Section 6.1.3. As a reference, the same sentences were synthesized with the model speaker database used for the prosody model training. The means and SDs of F_0 and duration of target units are shown in Table 7 under the column, "Target" and those of F_0 and duration of selected units are shown under the column, "Selected". The column, "Difference" shows the mean and SD of the difference between target and selected units ("Target" – "Selected") followed by "RMS error". Since the current implementation of CHATR does not predict RMS energy, comparison was not

Table 7
Target and selected units for speech synthesized by proposed method

Speech database		F_0 (Hz)			
		Target (mean ± SD)	Selected	Difference	RMS error
Female	Anger	259.9 ± 46.8	264.2 ± 44.3	−4.3 ± 38.9	39.1
	Joy	254.7 ± 42.3	253.7 ± 41.3	1.0 ± 38.0	38.0
	Sadness	238.9 ± 31.0	235.6 ± 25.1	3.3 ± 27.8	27.9
Male	Anger	160.6 ± 36.7	163.6 ± 35.1	−3.0 ± 47.0	47.0
	Joy	152.2 ± 32.9	148.4 ± 29.2	3.7 ± 38.07	38.1
	Sadness	128.5 ± 22.6	124.4 ± 13.0	4.0 ± 20.7	21.1
Model speaker		124.8 ± 23.6	121.6 ± 29.0	3.2 ± 28.6	28.7
		Duration (ms)			
Female	Anger	95.6 ± 39.5	74.3 ± 34.1	21.7 ± 25.5	33.4
	Joy	93.6 ± 40.3	74.3 ± 34.1	19.4 ± 26.9	33.1
	Sadness	107.3 ± 45.7	84.3 ± 40.9	23.0 ± 29.7	37.5
Male	Anger	85.9 ± 34.0	69.9 ± 31.8	16.0 ± 25.1	29.8
	Joy	82.6 ± 32.3	67.5 ± 32.0	15.1 ± 26.2	30.2
	Sadness	81.7 ± 27.4	67.5 ± 26.9	14.3 ± 18.5	23.4
Model speaker		74.4 ± 34.0	72.9 ± 32.3	1.5 ± 18.8	18.8

Model speaker: Speaker database used for prosody prediction training.

possible for RMS energy. For reference, mean and SD of RMS energy of selected units for each female speech corpus were anger: 56.5 ± 7.7 dB, joy: 58.1 ± 6.3 dB, and sadness: 55.1 ± 7.2 dB, and those for each male speech corpus were anger: 48.8 ± 6.1 dB, joy: 56.8 ± 6.3 dB, and sadness: 52.8 ± 7.8 dB.

As can be seen in Table 7, the means of differences between target and selected F_0s are very small for all emotions. As for SDs, when the SDs of target, selected, and the difference between them are the same, the correlation coefficients between target and selected units are 0.5. When SD of difference is smaller than those of target and selected units, there would be a stronger correlation between the two than when all three have equal values. Our data shows that SDs of difference of both F_0 and duration of all female speech and those of duration of all male speech are smaller than those of target and selected units. This then indicates a stronger correlation between targets and the selected for those data. RMS errors of F_0 and duration of speech synthesized with each emotional corpus are comparable to those of speech synthesized with the model speaker's corpus. This implies that the emotional corpora created in this study have sufficient variation of units to be used as a source database for CHATR synthesis.

We then observed differences between target and selected trajectories in more detail using the same speech samples of the female speaker shown in Section 3.3: Speech with the textual content of "Yoru, aniwa okiru mitaida (My brother seems to wake up at night)". As a reference, the trajectory of speech synthesized with her neutral corpus (a speech corpus of a reading of ATR 525 sentences) is shown as in Section 3.3. Fig. 11(a) shows target and selected F_0 trajectories and (b) shows target and selected duration per unit. By illustrating this way, the clear correlation is confirmed at a glance. Mean and SD of F_0 and duration of each corpus are shown on the right hand side of each graph.

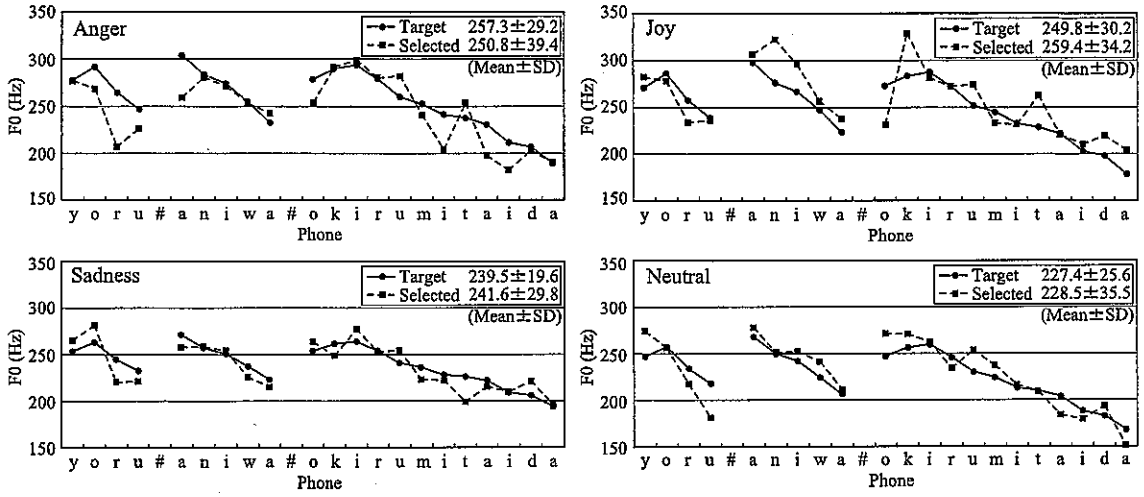
When compared with an acoustic profile of each corpus, both F_0 and duration values of selected units show clear correlations with the results of a prosodic analysis of the created corpora. Acoustic profiles of the emotional corpora are listed in Table 3 in Section 5.2.1 and those of neutral corpus are: F_0: 228 ± 46.7 Hz, duration: 75.1 ± 33.3 ms, and RMS energy: 57.6 ± 6.8 dB. These results are well predictable since the predicted prosodic patterns are mapped according to normalized values of the source database as described in Section 3.3. The result of the analysis in this section verifies that our method generates similar target F_0 and duration patterns for all emotions and that the entire pattern is either compressed or expanded, shifted to either a region of smaller values or larger values.

6.1.2. More careful observation at selected units

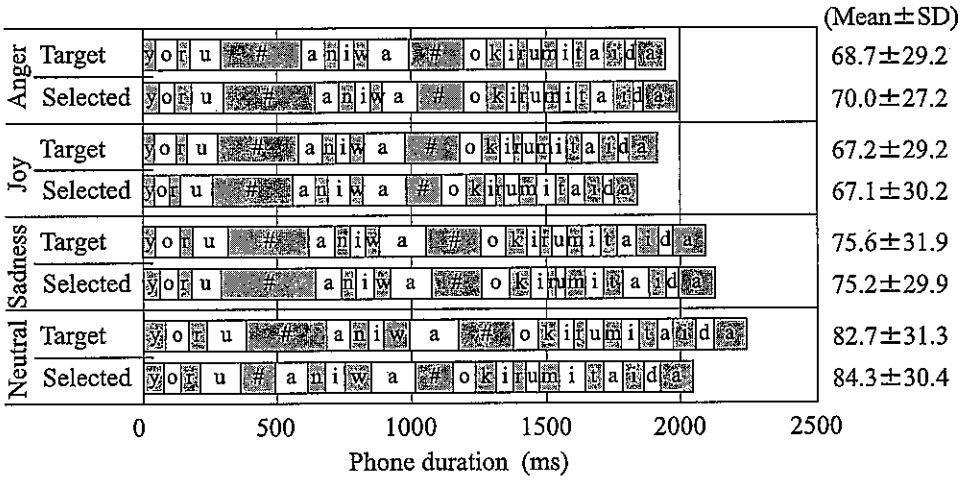
Observing the synthesized speech by selected units in comparison with target units, there were several incorrect positioning of the nuclear accents (e.g., giving the fourth mora an accent instead of the second in "obaasan", the third instead of the first for "aruhi" cf. Table 8). Since the prosody for these examples is correctly predicted by the prosody prediction model, the possible cause of incorrect positioning of accents is that several of the units selected were not close enough to their targets. This is largely attributed to a lack of suitable units in the source database. Also some incorrect positioning due to prosody prediction errors was found (e.g., giving second mora an accent instead of the first in "shisutemu (system)". These errors might influence the intelligibility of the synthesized speech. Therefore, some perceptual confusion may have resulted from the incorrect positioning of the lexical accents upon synthesis. However, we have not found concrete example yet and the nature and extent of such confusions is a topic for future work.

6.1.3. Emotion identification text

Subjects were asked to choose the emotion type of the synthesized speech stimuli from among anger, joy, and sadness as they were in the forced choice test performed with the recorded speech (cf. Section 5.1.2). The purpose of this experiment was also the same: to see if the subjects could differentiate the three kinds of emotional speech synthesized by our method.



(a) Target and selected F0 per phone



(b) Target and selected duration per phone

Fig. 11. F_0 and segmental duration of target and selected units for "Yoru, aniwa okiru mitaida (My brother seems to wake up at night)" using four kinds of corpus: anger, joy, sadness, and neutral of the female speaker. In (b) "#" indicates *pause* between phrases. Mean and SD of duration are calculated excluding pauses.

Subjects consisted of 18 university students (10 males, 8 females). The speech stimuli were the ones listed in Table 8 used for prosodic parameter analysis in Section 6.1.1 (5 sentences × 3 emotions of a male and a female speaker, which totals to 30 speech samples). These sentences were chosen from the CHATR demonstration materials. The selection criterion for these sentences was "No emotional modifiers were used".

Emotion identification rates for female speech are shown in Fig. 12 with anger: 60.0% joy: 51.1%, and sadness: 82.2%. Those for male speech are shown in Fig. 13 with anger: 51.1%, joy: 52.2%, and sadness: 74.4%. By applying a null hypothesis for the identification rate of each emotion (33.3% chance level), a test of the difference between the means of the corresponding binomial distribution rejected this hypothesis for all of the emotional

Table 8
Sentences used for emotion identification

Sentence ID	Sentence
1	CHATRwa iroirona koede shaberu kotono dekiru atarashii onseigoseino shisutemudesu. (CHATR is a new speech synthesizer that can speak in various voices.)
2	Ah, tsukareta. (Uh, I am tired.)
3	Aree, kaze hiita mitai. (Oops, I think I caught a cold.)
4	"Omae, Ningenwa kaoya naide." to iu tokorowo, "Omaeno kaowa ningenya naide." to itte shimota. Ahoyanaa. (Instead of saying, "Hey you, you cannot judge a person from his face," I said, "Your face is not the face of a human being." I made fool out of myself.)
5	Aruhi Obaasanga kawade sentakuwo shiteiruto, kawakamikara ookina momoga donburakodonburakoto nagarete kimashita. (One day, when an old woman was washing clothes by the river, there came a huge big peach splashing from the upper reaches of the river.)

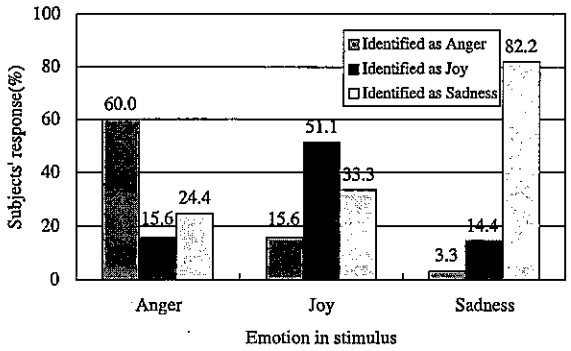


Fig. 12. Types of emotion identified for speech stimuli synthesized with emotional speech database of female speaker.

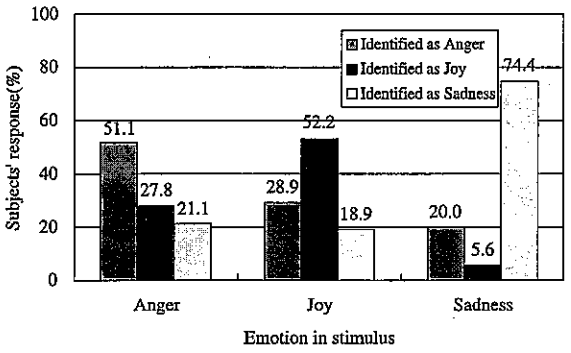


Fig. 13. Types of emotion identified for speech stimuli synthesized with emotional speech database of male speaker.

synthesized speech, at $p < 0.01$ (cf. Section 5.1.2). This result shows that the intended emotion was recognized at an incidence statistically greater than chance. The same statistical test also showed that the means of binomial distribution of the highest candidates were significantly different at $p < 0.05$ (and thus preferred) compared with the second highest candidates even though the latter were also higher than chance level (e.g. the female's joy speech identified as sadness).

For the female speaker, as with the identification test performed for her recorded speech (cf. Section 5.1.2), an additional question asked whether the emotion type of the speech stimuli could be judged from the textual content. This time, only 13% of the total speech stimuli were marked as such, whereas 47% were marked for the recorded speech. The lower rates here and the high rates in the identification test imply that subjects judged the emotion types not only from the textual content but also from the acoustic information in the speech stimuli.

For both speakers, speech expressing sadness was most successfully identified. The results using female synthesized speech correspond to those of the test using her recorded speech, and confusion of anger for sadness, and of joy for sadness, appears more noticeable. As for the male synthesized speech, anger and joy were shown to be confusion pairs. The confusion between anger and joy of the male synthesized speech might be attributed to their means and SDs of F_0 being closer, leaving sad speech distinguished as shown in Table 7 (anger: 160.6 ± 36.7 Hz, joy: 152.2 ± 32.9 Hz, sadness: 128.5 ± 22.6 Hz). The confusion of joy for sadness of the female synthesized speech cannot be explained from the results of the prosodic analysis. Spectral information might relate to this confu-

sion, but this is also difficult to explain from our formant data.

6.2. Evaluation towards practical use

To be used as a practical communication aid, it is important that the speech synthesized by our method be intelligible and gives the users and interlocutors a favorable impression when perceived. This section reports on the results of perceptual experiments evaluating those factors.

6.2.1. Intelligibility evaluation test

Subjects were 12 university students (9 males, 3 females) who transcribed the words they perceived from the speech stimuli synthesized by our method and by a commercial system for comparison. The objective of this experiment was to see if the proposed method could show intelligibility as high as a system in actual use. The commercial system for comparison is a CV + VC concatenative synthesis system (where C stands for consonant and V for vowel) with a signal processing module released in June 1997. It works on MS Windows 95 and higher, and its LSI version is implemented in a communication aid in common use in Japan.

The usual approach to evaluate the intelligibility of speech of a certain length is to administer a comprehension quiz after a listening session (Kitahara et al., 1987; JEIDA, 2000). However, with this method it is difficult to judge whether a subjects' incorrect answers are due to the synthesized speech being unintelligible or due to the content being incomprehensible. Therefore, we measured intelligibility by way of a dictation test where subjects were asked to write down the exact words on an answer sheet. The score for each sentence was calculated by summing up the number of correctly written *bunsetsu*. A bunsetsu is a syntactic unit in Japanese that consists of one content word (a word such as a noun or a verb that imparts meaning by itself) on the left, and one or more function morphemes (such as an auxiliary verb or a prepositional phrase) on the right. This corresponds to a phrasal unit such as a noun phrase or a verb phrase in English.

Prior to the intelligibility test using our method and the commercial system, we performed a similar kind of test with our method using all six speech variations (anger, joy, sadness of the male and the female). The aim of doing so is to reduce the number of speech stimuli in the test using both systems. Eighteen subjects listened to two context-neutral sentences and wrote down what they heard. The results of bunsetsu scoring were analyzed by a t-test at $p < 0.05$, and no significant difference was found among the six kinds of speech.

Sentences used in the test with both systems contained four to seven bunsetsus, as shown in Table 9 (for ease in identifying bunsetsus, slashes are placed between them). All sentences were synthesized by both systems. Three types of sentence were prepared: (1) two sentences with numbers, (2) two sentences with uncommon noun collocations, and (3) two sentences with emotional expressions. These sentence variations were selected on the basis of the following assumptions: (1) numbers are important and are often misinterpreted, which can lead to misunderstanding between interlocutors, (2) examining intelligibility with pairs of uncommon collocations can help to avoid guessing, and (3) sentences with emotional expressions are often used in the target system of the proposed method.

In order to avoid the semantic influence of a sentence on a speech stimulus, subjects were divided into two groups (A, B) of six people and sentences were also divided into two groups (X, Y) of three sentences. The sentence-stimulus combination for subject group A was reversed for subject group B, as shown in the two right-most columns in Table 9. The selection between female and male speaker was randomly decided for our method, and then the speaker setting of the commercial system was set to the corresponding gender.

The results for each sentence are shown in Fig. 14. While both systems yield high intelligibility, the average intelligibility rate for all six sentences for our system (92%) was higher than that for the commercial synthesizer (81.9%). When sentence types are compared, for both systems, the score was the lowest for sentences with uncommon collocations (79% for our system and 61% for the commercial system) and the highest for sentences including numbers (99% for both systems). The result of a t-test with a significance level at

Table 9
Sentences used for intelligibility test

Sentence			Number of syllables	Sentence	Speech stimulus	
ID	Type	Group			Group A	Group B
1	U	X	4	Akira-kunwa/yamazakurato/kakinabeto/iimashita. (Akira-kun said a mountain cherry tree and an oyster pot).	Emo-M Sad	Com-M
2	N	X	7	Shidoniito/Tokyono/jisawa/ichijikandesunode/imawa/asano/8ji15fundesu. (The time difference between Sydney and Tokyo is an hour so it is now 8:15 in the morning.)	Emo-F Ang	Com-F
3	E	X	5	Wa-i,/yatto/kurundane./Ganbattekite/yokatta. (Wow, so he is finally comming! I am glad I have been trying so hard.)	Emo-M Joy	Com-M
4	U	Y	4	Yamamoto-sanwa/Kaichudentoto/kureyonto/iimashita. (Ms. Yamamoto said a torch and a crayon.)	Com-F	Emo-F Joy
5	N	Y	5	8gatsu15nichino/nichiyoubikara/8gatsu/16nichino/getsuyoubi,/ippakufutsukadesu. (From Sunday, August, 15 from Monday, August, 16, a one night two days trip.)	Com-M	Emo-M Ang
6	E	Y	5	Ah,/tsukareta./Hajimetenanode/totemo/fuandesu. (Oh, I am tired. I am very worried since it is my first time.)	Com-F	Emo-F Sad

U: a sentence with uncommon noun collocations, N: a sentence with numbers, E: a sentence with emotional expressions. Emo: synthesised speech by proposed method, Com: synthesised speech by a commercial system, M: male, F: female. Ang: Anger, Sad: Sadness. Slashes are placed in between bunsetsus.

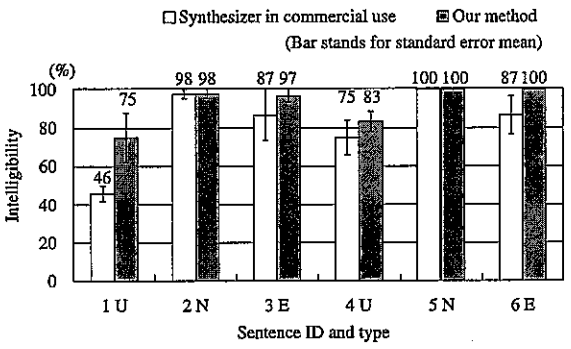


Fig. 14. Intelligibility per sentence of our system compared with that of a commercial system. For sentence type, see text and Table 9.

$p < 0.05$ showed that there was no significant difference between the two synthesizers. This result is sufficient to confirm that our system synthesizes sufficiently intelligible speech equivalent to speech synthesized by the commercial system, and, therefore, it can be used in practical situation. Along with the positive results of emotion identification tests and the subjective impression tests described in the next section, it can be said that to a certain extent our system can offer an added value of expressiveness.

6.2.2. Evaluation of subjects' impressions

In this experiment, we investigated whether the speech synthesized by our method gives a favorable impression. The objective of this evaluation is to ascertain a rough idea of the 'likes' and 'dislikes' of subjects when they listen to the synthesized speech. Again, a comparative approach was taken. All sentences were synthesized by our method and by a communication aid released in May 1999. It uses Ricoh RL5S850, a CV + VC concatenative synthesis LSI released in April 1998 whose base system is the same as the one used in the intelligibility test in Section 6.2.1.

Subjects were 143 university students (69 males, 74 females). For speech stimuli, emotive sentences were synthesized with speech databases with corresponding emotion, assuming the actual usage of the proposed system. As shown in Table 10, for our method, three sentences that each expressed anger, joy, and sadness, and a short passage of six sentences – two sentences expressing anger, two joy, and two sadness – were used for synthesis with a source database of corresponding emotions. Taking the same procedure as in the intelligibility test (cf. Section 6.2.1), the selection between female and male speaker was randomly decided for our method, and then the speaker setting of the com-

Table 10
Sentences used for subjective evaluation

Speech stimulus			Sentence
ID	Our method	Commercial system	
1	Male Joy	Male	Wa-i, yatto kurundane. Ganbattekite yokatta. (Wow, so he is finally coming! I am glad I have been trying so hard.)
2	Male Ang	Male	Mattaku dou kangaetemo abunainoni darega sonna kotowo shitanda. (It is dangerous and there is no question about it, but who in the earth did such a thing.)
3	Female Sad	Female	Ah, tsukareta. Yorumo nemurazuni ewo kakitsuzukete kigatsuitara asani natte itanda. (Oh, I am tired. I kept on drawing all through the night and I did not realize what time it was until the morning.)
4	Male Joy	Male	Bokuwa shinshinkieino sakkyokuka desu. (I am a young and promising composer.)
	Male Sad	Male	Mada kakedashinande minna bokuwo bakani surundesukedone. (I am still a beginner, so people would look down on me.)
	Male Ang	Male	Imani urekkoni natte miserukara miteroyo. (One day I will become popular so watch it!)
	Female Joy	Female	Itsumo arigatou gozaimasu. Watashiwa sokono chuugakuno sannenseidesu. (Thank you very much. I am in the 3rd grade of the junior high school over there.)
	Female Sad	Female	Mada yaritai kotoga yoku wakaranaindesukedone. (I do not know what I want to be yet.)
	Female Ang	Female	Demo dakaratte minnade noromatte iundesu. Mattaku atamani kimasu. (But that does not mean they can call me dopy. I am really mad.)

Ang: Anger, Joy: Joy, Sad: Sadness.

mercial system was set to the corresponding gender in this test. Subjects listened to the speech stimuli through speakers (SONY SRS A21) in a classroom and were asked to rate each speech stimulus on a 5-point scale (5: excellent, 1: very poor) to indicate their overall impression. Mean opinion score (MOS) and SD are shown in Fig. 15. MOS was higher for our method (overall MOS $\pm$ SD $= 2.8 \pm 0.98$) than for the commercial system (1.9 ± 1.01). The result was analyzed by t-test at $p < 0.05$, and confirmed a significant difference between the two systems.

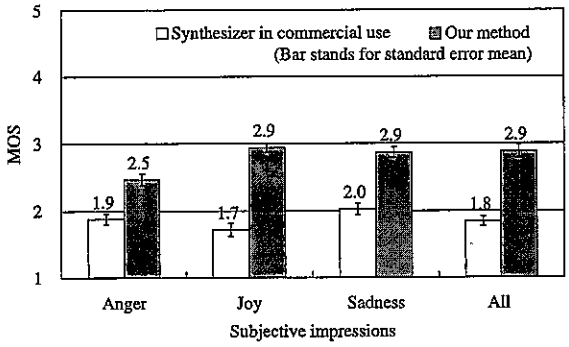


Fig. 15. Subjective impressions of university students.

6.2.3. Evaluation by target users

We individually interviewed five patients who were suffering from Amyotropic Lateral Sclerosis (ALS), aged 46–61 years old (3 males and 2 females). ALS is a motor neurone disease, and at the time of the interviews, patients had disabilities due to their bodies being paralyzed. Among them, three had lost their speaking ability due to tracheotomies. The objective of having these target users evaluate our method was simply to get their general impressions.

A forced choice identification test, an intelligibility test, and an overall impression test were performed. In order to minimize the patients' workload, the emotive texts used in the subjective impression test (Table 10), which could best introduce the characteristics of our method, were used for all three experiments. Loudspeakers (SONY SRS A21) were used for listening to the speech. For each sentence in the short passage (cf. No. 4 of Table 10), speech stimuli were synthesized and presented one-by-one. Since subjects had difficulty in writing, the interviewer (the first author) asked questions and transcribed their answers. For the intelligibility test, subjects were asked to indicate which words they could not understand by

pointing to the target word in the sentence written on the paper after listening to the synthesized speech.

The identification rates for speech stimuli synthesized in both male and female voice were quite high: anger: 93.3%, joy: 66.6%, and sadness: 86.6%. This is mainly due to intrinsic suggested emotional content paired with appropriate emotion. A test of the difference between the means of binomial distributions at $p < 0.01$ shows that the intended emotion was identified at an incidence significantly greater than chance. As for intelligibility, 91.7% correctness was obtained for our method and 85.2% for the commercial communication aid. The result of a t-test with significance level at $p < 0.05$ showed that there was no significant difference in intelligibility between the two synthesizers. This result has a good indication as seen in Section 6.2.1. MOS was higher for our method (overall MOS $\pm$ SD $= 3.4 \pm 1.19$) than for the communication aid (2.2 ± 1.00). Although it was difficult to draw any quantitative assessment from the perceptual tests of only five subjects, we can say from the above results that the system can be considered as suitable for practical use.

6.2.4. Identification test adding neutral speech

From the interview with target users, we found that adding a source database of neutral speech would be required in practical use. Emotionally neutral is difficult to define, but as a trial, we recorded a reading of the ATR 525 sentences by the same female speaker in her normal reading manner. This corpus was recorded with an interval of four years after the recording of emotional speech corpora at a different location and with different recording equipment. The acoustic profile of this corpus is shown in Section 6.1.1 but voice quality including F_0 might have changed from the time the emotional speech corpora were recorded. The male speaker corpus is under development.

A forced choice test was conducted to see if subjects could identify emotion types from four kinds of synthesized speech: anger, joy, sadness, and neutral. Speech stimuli were presented via the Internet. Subjects were 30 individuals (16 males, 14 females) who listened with headsets. Other conditions for the experiment were described in the introductory paragraph of Section 6. Since some of them participated in previous experiments, four new sentences, carefully designed to be context-neutral, were used as shown in Table 11. Five choices (anger, joy, sadness, neutral, cannot tell) were given as possible responses.

As shown in Fig. 16, the identification rates were anger: 51.7%, joy: 36.7%, sadness: 79.2%, and neutral: 51.7%. The rate of joy speech identified as joy (36.7%) is above chance level (20%) but lower than the rate of joy identified as neutral (44.2%). Other noticeable confusions are anger identified as neutral by 24.2% and neutral identified as sadness by 30.8%.

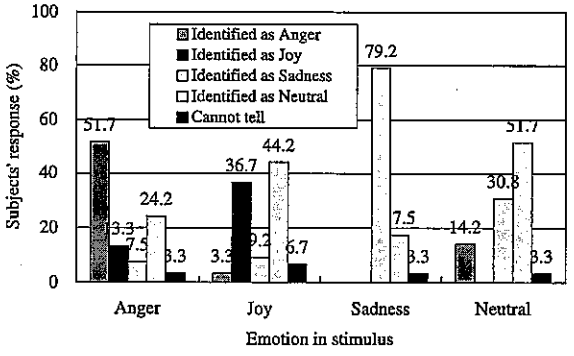


Fig. 16. Results of identification test among four styles of speech synthesized with female emotional speech database.

Table 11
Context-neutral sentences used for identification test

Sentence ID	Sentence
1	Asokoni keiyakushoga arimasu. (There is the contract form.)
2	Ashitawa minnade Ooyamakara kaerimasu. (We will return from Ooyama tomorrow.)
3	Chuumon shitekara isshuukande seihinga todokimashita. (The product reached us 1 week later we ordered.)
4	Yukikowa watashino anedesu. (Yukiko is my older sister.)

Setting a chance level as 20%, the result of the test of the difference between the means of binomial distributions indicated that the highest candidate was identified at an incidence greater than chance at a significance level of $p < 0.01$. Also the same statistical test showed that the means of binominal distributions of the highest candidate were significantly different at $p < 0.05$ and, therefore, that the highest candidate was preferred over the second highest candidate even though the latter also had higher rates than chance level. From these results, it can be said that the synthesized speech expressing anger, sadness, and neutral were identified at an incidence statistically greater than chance. However, as discussed, joy speech was not successfully identified as joy since the identification rate of joy speech identified as neutral was higher than joy speech identified as joy. It is understandable that sad speech might have been identified as neutral, since F_0 and duration of corresponding corpora are alike. F_0 range of neutral is closer to joy but also closer to anger compared to the sad corpus. It might be possible to consider that two corpora are alike in some other acoustic profiles. Since the neutral speech was selected either as the highest or the second highest candidates in all emotions, some subjects may have set a normative or a default speech to neutral when judgment was difficult during the identification test.

We further performed an identification test using emotive sentences to confirm the validity of the joy corpus in actual use. Subjects were 30 individuals (among them, five males and five females were replaced from the previous test using neutral sentences). The same test was performed for the anger corpus as well. Four emotive text materials (shown in Table 12) were newly prepared (two sentences expressing anger and two expressing joy), and these were synthesized using all four source databases (anger, joy, sadness, and neutral) of the female speaker. Consequently, the speech stimuli produced with source databases other than that of the intended emotion were used as distracters. Five choices ("anger", "joy", "sadness", "neutral", and "cannot tell") were given as possible responses. The result is as follows: for anger sentences synthesized with the anger corpus, 90.1% selected anger, and for joy sentences synthesized with joy corpus, 80.5% selected joy. Both joy and anger were identifiable at an incidence significantly greater than chance. For both angry and joyful sentences synthesized with source databases other than that of the intended emotion, the probabilities of them being selected as anger or joy were much smaller. The results of all combinations are shown in Table 13.

Table 12
Emotive sentences used in identification test for joy and anger

Sentence ID	Emotion	Sentence
1	Joy	Arigatou gozaimasu. (Thank you very much.)
2	Joy	Waai, tsuini kansei shitandane. (Wow, I finally accomplished it.)
3	Anger	Iikagenni mewosamase. (Why don't you wake up to reality.)
4	Anger	Ikinari nante haradatashii mattaku. (He was so sudden. I am so offended, yes absolutely.)

Table 13
Types of emotion identified from four styles of synthesized speech using emotive sentences (anger, joy, sadness and neutral)

Sentence	Emotion in stimulus	Subjects' response (%)				
		Anger	Joy	Neutral	Sadness	Cannot tell
Anger	Anger	90.0	1.7	3.3	0.0	5.0
	Joy	31.7	23.3	38.3	1.7	5.0
	Neutral	5.0	1.7	28.3	60.0	5.0
	Sadness	26.7	0.0	46.7	13.3	13.3
Joy	Anger	21.7	33.3	30.0	8.3	6.7
	Joy	0.0	81.7	16.7	0.0	1.7
	Neutral	0.0	10.0	26.7	61.7	1.7
	Sadness	3.3	3.3	60.0	28.3	5.0

7. Implementing a TTS system with emotion

To support the immediate needs of our target users, we implemented Chatako, a PC-based TTS system with emotion, incorporating the proposed method. Chatako runs on Microsoft Windows 95 and higher using CHATR98 and can synthesize the emotional speech of anger, joy, and sadness. Currently, the female neutral speech is incorporated as neutral speech and that of the male speaker is underway. A target user with physical disabilities can use this system with an additional input device developed for his or her disabilities.

7.1. Design principles

This system is designed for people who have lost both their ability to speak and the means of expressing their emotions, either vocally or even physically, due to paralysis. Selectable voices are one male and one female voice, and selectable emotion types are anger, joy, and sadness. A unique feature of this system is that it can change the emotion types of the synthesized speech while synthesizing, as if the speaker changes his/her speaking style during an utterance. This function is realized by switching between the source databases, and users can do this by a simple procedure described in Section 7.3.

7.2. Chatako's graphical user interface

As shown in Fig. 17, Chatako's GUI is written in Tcl/Tk 8.3 and is equipped with a text window, selectable speaker icons, selectable emotion icons, and command buttons. The text window is located at the center, and the speaker icons ("male" and "female") and emotion icons (anger, joy, and sadness) are located to the right of the text window. Beneath the text window are command buttons. In this prototype, the available commands are "open", "whole synthesis", "part synthesis", "save speech", "save sentence", "window clear", and "exit", shown in left-to-right order in Fig. 17. In this way, all of the selections are displayed in the window so that the user is only required to make a single click for each command, unlike pull down menus that require several mouse clicks.

7.3. Speaker-emotion selection

When the user selects a speaker by clicking either the "male" or the "female" icon, emotion icons of that speaker appear beneath the speaker

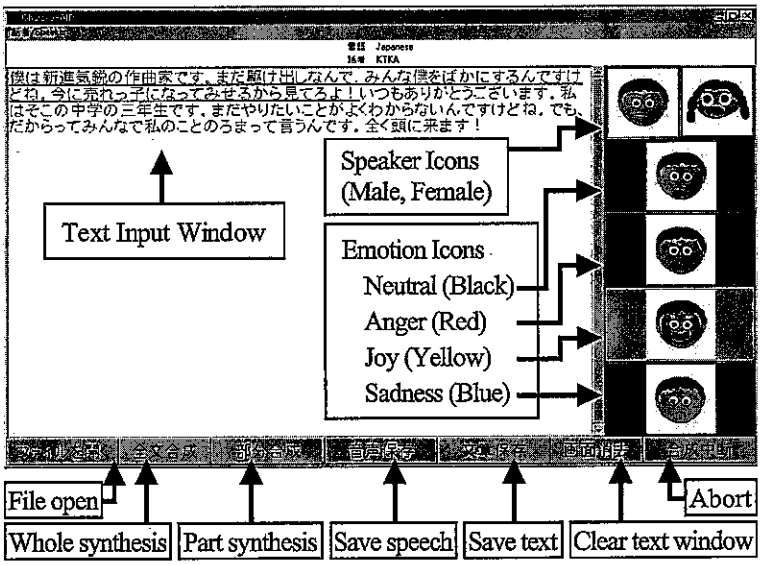


Fig. 17. Chatako's window layout. Text in the text window is the short passage in Table 10 (No. 4) written in Japanese orthography.

Table 14
Example of stored tagged text

Tagged text	Translation
TAGMale: joy bokuwa shinshin kieino sakkyokuka desu.	I am a young and promising composer.
TAGMale: sadness mada kakedashi nandesukedone.	I am still a beginner, so people would look down on me.
TAGFemale: anger mattaku atamani kimasu.	I am really mad.

icons, and then the user selects the emotion type. The user then types the text into the text window. By doing this, the system automatically generates a text file with speaker-emotion tags and passes that information to the synthesis module so that the synthesized speech will be produced with the corresponding database when the user selects the "synthesize" command. Table 14 shows an example of saved tagged text.

When the user wishes to change a speaker or emotion in the middle of typing, he or she can do this by re-selecting the speaker and emotion icons and starting to type again. Speaker-emotion information once tagged to text can be changed by partially marking the portion of the text that the user wishes to change and reselecting the speaker and/or emotion. A text file with speaker-emotion information tags is stored when user selects the "save text" command and can be opened with the "file open" command. Speaker information is distinctively displayed by using underlined fonts for the male speaker and plain fonts for the female speaker (this choice is selectable). Emotion types are distinguished by font colors: red for anger, yellowish green for joy, blue for sadness, and black for neutral state.

7.4. Commands for synthesis

Three ways are provided to call the synthesis module. The first is to synthesize the entire message, the second is to synthesize a part of the message, and the last is to synthesize line-by-line. The first function is executed by the "whole synthesis" command, which synthesizes all of the sentences in the text window. This function is convenient when the user wishes to synthesize a large amount of text. Such a need occurs when the user wants to give a talk or to participate in a conference. The users can make use of the standard copy and paste commands of MS Windows to copy text from other text files to Chatako's text window. This second function is executed by the "part synthesis" command. This function synthesizes only the marked portion and is convenient when the user wishes to repeat the same utterance a number of times, for instance, when the interlocutor does not hear the synthesized speech correctly or when the user wishes to express appreciation repeatedly to either the same interlocutor or to someone else who comes right after the original interlocutor. The last function is executed by hitting a carriage return. This function enables the user to synthesize texts from the point immediately after the previously synthesized part to where the cursor is currently located. This function is derived from a simulation of the real-time conversation between the user and the interlocutor.

8. Discussion

In this study, we statistically analyzed a large amount of emotional speech data to confirm the finding of earlier studies that F_0 and duration were important acoustic correlates of emotion. Some confusion of emotion types observed in the emotion identification test of speech synthesized by our method could not be clearly explained by the prosodic parameters measured.

While emotional cues are certainly included in prosodic information, the formant analysis performed in this study indicated the possibility that emotional cues may also be included in spectral information. Further research should be carried out to identify acoustic features in both suprasegmental and segmental information that are relevant for specifying a particular emotion in a parametric way. By doing so, those features could

be included as selection criteria for unit selection synthesis. This addition would lead to the enrichment of the source database and it would potentially allow us to synthesize emotional speech from any natural recording of a single speaker, on the assumption that human speech contains various emotional expressions when naturally produced over a relatively long period of time.

9. Conclusions

This paper reported a new approach to synthesizing emotional speech using the corpus-based concatenative speech synthesis system (ATR CHATR) with speech corpora of emotional speech. Three kinds of emotional speech (anger, joy, and sadness) were created from a male and a female speaker of Japanese for ATR's CHATR. The advantage of this method is that any emotion or speaking style can be synthesized by creating the appropriate source database. The results of perceptual experiments confirmed that our method can synthesize recognizably emotional speech that provides high intelligibility and gives a favorable impression. A workable PC-based TTS system incorporating the proposed method was developed for use by nonspeaking individuals. Future directions include identifying acoustic features that are relevant for specifying emotions in a parametric way, so that those features could be used in defining emotion-dependent labels, improving the unit selection algorithm, and building emotion-dependent prosodic models.

Acknowledgements

This research was partially supported by JST CREST. The authors express their appreciation to Dr. Parham Mokhtari of JST CREST for preparing the steady state vowels and the formant data. The authors also thank Mr. Shinnichi Yamaguchi, of Fukuoka, Japan, Dr. Soichiro Iga of Ricoh Co., Ltd., Mr. Marc Schröder of the University of the Saarland, and Mr. Kazuya Takahashi of Keio University for valuable discussions. The authors would further like to thank target users, students, and colleagues who participated in the experiments.

References

Abe, M., Sagisaka, Y., Umeda, T., Kuwabara, H., 1990. ATR Technical Report TR-I-0166 Speech Database User's Manual. ATR Interpreting Telephony Research Lab.

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70 (3), 614–636.

Black, A., Hunt, A., 1996. Generating F_0 contours from ToBI labels using linear regression. In: Proceedings of ICSLP 96, Philadelphia, USA, Vol. 3, pp. 1385–1388.

Bunnell, H.T., Hoskins, S.R., 1998. Prosodic vs segmental contributions to naturalness in a diphone synthesizer. In: Proceedings of ICSLP 98, Sydney, Australia, Vol. 5, pp. 1723–1726.

Cahn, J.E., 1989. Generation of affect in synthesized speech. In: Proceedings of the 1989 Conference of the American Voice I/O Society, pp. 251–256.

Campbell, W.N., 1996. Autolabelling Japanese TOBI. In: Proceedings of ICSLP 96, Philadelphia, USA, pp. 2399–2402.

Campbell, W.N., 1997a. Processing a speech corpus for CHATR synthesis. In: Proceedings of International Conference on Speech Processing (ICSP'97), Seoul, Korea, pp. 183–186.

Campbell, W.N., 1997b. Synthesizing spontaneous speech. In: Sagisaka, Y., Campbell, N., Higuchi, N. (Eds.), Computing Prosody. Springer, NY, pp. 165–186.

Campbell, W.N., Black, A., 1997. Prosody and the selection of source units for concatenative synthesis. In: van Santen, J., Sproat, R., Olive, J., Hirschberg, J. (Eds.), Progress in Speech Synthesis. Springer Verlag, NY, pp. 279–292.

Carlson, R., Granstrom, G., Nord, L., 1992. Experiments with emotive speech, acted utterances and synthesized replicas. In: Proceedings of ICSLP 92, Banff, Canada. Vol. 2, pp. 671–674.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. Emotion recognition in human–computer interaction. IEEE Signal Processing Magazine 18 (1), 32–80.

Entropic Research Laboratory, Inc., 1996. ESPS Programs A–L.

Davitz, J.R., 1964a. A review of research concerned with facial and vocal expressions of emotion. In: Davitz, J.R. (Ed.), The Communication of Emotional Meaning. McGraw-Hill, NY, pp. 13–29.

Davitz, J.R., 1964b. Auditory correlates of vocal expressions of emotional meanings. In: Davitz, J.R. (Ed.), The Communication of Emotional Meaning. McGraw-Hill, NY, pp. 101–112.

Fairbanks, G., Pronovost, W., 1939. An experimental study of the pitch characteristics of the voice during the expression of emotion. Speech Monographs 6, 87–104.

Guerrero, L.K., Andersen, P.A., Trost, M.R., 1998. Communication and emotion: basic concepts and approaches. In:

Andersen, P.A., Guerrero, L.K. (Eds.), Handbook of Communication and Emotion: Research, Theory, Applications and Contexts. Academic Press, San Diego, pp. 3–27.

Ichikawa, A., Nakayama, T., Nakata, K., 1967. Experimental consideration on the naturalness of the synthesized speech. In: Proceedings of Acoustic Society of Japan Fall Meeting, pp. 95–96 (in Japanese).

Iida, A., Iga, S., Higuchi, F., Campbell, N., Yasumura, M., 1998. Acoustic nature and perceptual testing of a corpus of emotional speech. In: Proceedings of ICSLP 98, Sydney, Australia, Vol. 4, pp. 1559–1592.

Iida, A., Iga, S., Higuchi, F., Campbell, N., Yasumura, M., 2000. A speech synthesis system with emotion for assisting communication. In: Proceedings of ISCA Workshop on Speech and Emotion, Belfast, UK, pp. 167–172.

Ito, K., 1986. A basic study on voice sound involving emotion (III). Ergonomics 22 (4), 211–217 (in Japanese).

JEIDA, 2000. The guideline of speech synthesis system performance evaluation methods. Japan Electronic Industry Development Association.

Kamimura, K., 1990. Ashitawo tsukuru – Sekizui sonshoushano seikatsuno kiroku. Miwashoten, Tokyo (in Japanese).

Katae, N., Kimura, S., 2000. An effect of voice quality and prosody control in emotional speech synthesis. In: Proceedings of Acoustic Society of Japan Fall Meeting, pp. 187–188 (in Japanese).

Keating, P.A., Huffman, M.K., 1984. Vowel variation in Japanese. Phonetica 41, 191–207.

Kitahara, Y., Tohkura, Y., 1992. Prosodic control to express emotions for man-machine speech interaction. Institute of Electronics, Information, and Communication Engineers (IEICE) Transactions of Fundamentals of Electronics E75-A (2), 155–163.

Kitahara, Y., Takeda, S., Ichikawa, A., Tohkura, Y., 1987. Role of prosody in cognitive process of spoken language. Journal of Interaction, Institute of Electronics, Information, and Communication Engineers (IEICE) D J70-D (11), 2095–2101 (in Japanese).

Makino, T., Arai, S., Ozawa, S., 1989. A method of vowel recognition in connected speech using the mutual relation of vowels. Institute of Electronics Information and Communication Engineers (IEICE) D-II J72-DII (6), 837–845 (in Japanese).

Mokhtari, P., Iida, A., Campbell, N., 2001. Some articulatory correlates of emotion variability in speech: a preliminary study on spoken Japanese vowels. In: Proceedings of the International Conference on Speech Processing (ICSP'01), Taejon, Korea, pp. 431–436.

Mozziconacci, S.J.L., 1998. Speech variability and emotion: production and perception. Ph.D. thesis, Technical University Eindhoven.

Murray, I.R., Arnott, J.L., 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. Journal of Acoustic Society of America 93 (2), 1097–1108.

Murray, I.R., Arnott, J.L., 1995. Implementation and testing of a system for producing emotion-by-rule in synthesized speech. Speech Communication 16, 369–390.

Murray, I.R., Arnott, J.L., Alm, N., Newell, A.F., 1991. A communication system for the disabled with emotional synthesized speech produced by rule. In: Proceedings of Eurospeech '91, Genova, Italy, pp. 311–314.

Murray, I.R., Edgington, M.D., Campion, D., Lynn, J., 2000. Rule-based emotion synthesis using concatenated speech. In: Proceedings of ISCA Workshop on Speech and Emotion, Belfast, UK, pp. 173–177.

Nagae, Y., 1998. Onsei ni fukumareru washa no kanjo no bunseki to ninshiki ni kansuru kenkyuu. Bachelor's thesis submitted to Utsuyomiya University (in Japanese).

Ohira, Y., 1995. Watashirashiku, ningenrashiku, Autobiography (in Japanese).

Russell, J.A., 1989. Measures of emotion. In: Plutchik, R., Kellerman, (Eds.), Emotion Theory, Research, and Experience, Academic Press, NY, Vol. 4, pp. 83–111.

Scherer, K.R., 1986. Vocal affect expression: a review and a model for future research. Psychological Bulletin 99 (2), 143–165.

Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck, T., 1991. Vocal cues in emotion encoding and decoding. Motivation and Emotion 15 (2), 123–148.

Shaver, P., Schwartz, J., Kirson, D., O'Connor, C., 1987. Emotion knowledge: Further exploration of a prototype approach. Journal of Personality and Social Psychology 52 (6), 1061–1086.

Takeda, S., Ishizuka, F., Hiramatsu, M., 2000. Power features of "anger" expressions in pseudo-conversational speech. In: Proceedings of Acoustic Society of Japan Fall Meeting, pp. 191–192 (in Japanese).

Todoroki, T., 1993. Kousai - Kagayaki tuzukeru tameni – KB mausude nyuuryokushita Kinzisutorofii seinenno kiroku (Self-published in Japanese).