INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 379–392, 2003

Speech Database Design for a Concatenative Text-to-Speech Synthesis System for Individuals with Communication Disorders

AKEMI IIDA
Keio Research Institute at SFC, Keio University; JST (Japan Science and Technology), CREST
akeiida@sfc.keio.ac.jp

NICK CAMPBELL
JST (Japan Science and Technology), CREST; ATR Human Information Sciences Research Laboratories

Abstract. ATR's CHATR is a corpus-based text-to-speech (TTS) synthesis system that selects concatenation units from a natural speech database. The system's approach enables us to create a voice output communication aid (VOCA) using the voices of individuals who are anticipating the loss of phonatory functions. The advantage of CHATR is that individuals can use their own voice for communication even after vocal loss. This paper reports on a case study of the development of a VOCA using recordings of Japanese read speech (i.e., oral reading) from an individual with amyotrophic lateral sclerosis (ALS). In addition to using the individual's speech, we designed a speech database that could reproduce the characteristics of natural utterances in both general and specific situations. We created three speech corpora in Japanese to synthesize ordinary daily speech (i.e., in a normal speaking style): (1) a phonetically balanced sentence set, to assure that the system was able to synthesize all speech sounds; (2) readings of manuscripts, written by the same individual, for synthesizing talks regularly given as a source of natural intonation, articulation and voice quality; and (3) words and short phrases, to provide daily vocabulary entries for reproducing natural utterances in predictable situations. By combining one or more corpora, we were able to create four kinds of source database for CHATR synthesis. Using each source database, we synthesized speech from six test sentences. We selected the source database to use by observing selected units of synthesized speech and by performing perceptual experiments where we presented the speech to 20 Japanese native speakers. Analyzing the results of both observations and evaluations, we selected a source database compiled from all corpora. Incorporating CHATR, the selected source database, and an input acceleration function, we developed a VOCA for the individual to use in his daily life. We also created emotional speech source databases designed for loading separately to the VOCA in addition to the compiled speech database.

Keywords: AAC, corpus-based TTS synthesis, speech corpus, communication disorder, VOCA

1. Introduction

There are many groups of people who are unable to communicate either vocally or physically but have unimpaired cognitive abilities. In this paper, we refer to such a state as a communication disorder. Common causes of these communication disorders include cerebral palsy, strokes, muscular dystrophy, and motor neuron diseases (MND) (National Institute of Neurological Disorders, 2001). Not only do these individuals not speak, but they also often suffer severe physical disorders, involving extreme difficulty in conveying intention to others as the symptoms progress. This catastrophic state, in some cases, deprives people of the will to live.

Augmentative and alternative communication (AAC) is an area of clinical practice that attempts to compensate for the impairment and disability of

individuals with severe communication disorders. The field makes use of both low-tech strategies, involving pointing to words and icons, and high-tech devices such as voice output communication aids (VOCA) (University of Nebraska-Lincoln, Aphasia Group, n.d.). In this section, we briefly describe some concepts and studies in AAC that are related to our study.

The first area to be introduced is the text corpus research. This area investigates words and phrases that are necessary for AAC device users and classifies them to two categories: a core (i.e., standard) and an extended (i.e., user) vocabulary. A core vocabulary includes words and phrases that are commonly used by target users and an extended vocabulary includes words and phrases that are used individually. Developing a dictionary component based on this classification allows users of AAC devices to select frequently used words with a few operations. Research in this area includes Beukelman et al. (1984) who analyzed words and phrases generated by five non-speaking adults using communication aids for everyday conversation. Another example is a research conducted by Yorkston et al. (1988) that compared and contrasted 11 standard vocabularies and 9 user vocabularies.

Another important area in AAC research is investigating functional methods that generate messages with as little physical action as possible. Vanderheiden and Kelso (1987) called these "acceleration techniques". Representative examples include abbreviation expansion, full sentence storage and retrieval, and predictive text input.

Voice output attracted attention in AAC as consumer electric devices became widely available. VOCA stored digitized speech for voice output in the early 1970s, but while the speech produced was of good quality, the range of vocabulary was limited.

In the 1980s, as systems that synthesized texts became available for practical use, text-to-speech (TTS) synthesis began to be used in VOCA. Until recently, most researchers used TTS synthesis systems within the limit of a synthesizers' performance (e.g., Cambridge Adaptive Communication, 2002; Hitachi Keiyo Engineering and Systems Ltd., n.d.). Formant and concatenative synthesizers with uniform-sized units were most commonly used. Speech produced by these synthesizers is intelligible (i.e. what synthesizers say is understandable), and many offer several different voices, ranging from child to adult in both male and female registers, e.g., DECtalk, the formant synthesizer (Conroy, 1986; Hallahan, 1996). There are, however, a number of problems to be resolved. Formant synthesizers use a non-human voicing source produced by an excitation signal that, to some, makes the synthesized speech sound robot-like and unnatural. In contrast, the waveform concatenation approach with uniform units uses the human voice as a source, which makes the synthesized speech sound more natural. However, the approach introduces some distortion when the speech signals are modified at unit concatenation. Recent attempts to adopt a natural speech corpus for unit selection in concatenative TTS have resulted in high quality natural speech using the human voice, but without applying the signal modification that produces distortion, if appropriate sub-word units are available in the corpus.

This paper reports on a case study to create a speech database for a concatenative TTS synthesis system by recording the read speech of a Japanese male individual with ALS (hereafter, the speaker) who is anticipating the loss of phonatory functions. Positive results of the study would suggest that other target users with similar symptoms (cf. Section 2) may be able to continue communicating, using their own voices, even after losing their natural speaking abilities. Further, the study aims to design a speech database that reproduces the speaker's natural utterances for both open domain (i.e., any text) and limited domain synthesis (i.e., words and/or phrases for selected situations). In this study, the intended situation is the daily interaction between the speaker and family members, caretakers, and friends.

The first step of our study was to prepare text corpora. In this respect, our study is similar to text corpus design in a traditional AAC scheme. The difference between our study and a traditional approach is that AAC researchers use text corpora only for deciding the vocabulary for a VOCA, while we used text to create speech corpora to improve the quality of synthesized speech.

The following section describes the target users of our system. Section 3 describes the CHATR system (Campbell and Black, 1997) designed at the Advanced Telecommunications Research Institute (ATR) – the speech synthesis system used in this study. Section 4 reports on how we created speech corpora and Section 5 describes evaluation procedures and results. Section 6 describes our development of a VOCA that we named Chatako-AID.

2. Target Users

Because working vocal functions are needed to create speech corpora for Chatako-AID, our target users are generally individuals with progressive diseases who can speak at the time of recording but who anticipates the loss of phonatory function. Examples of these diseases are neurological disorders such as motor neurone disease (MND) and muscular dystrophy that cause weakening of respiratory muscles. Throat cancer may also take away phonatory functions from patients, but this disease does not affect motor neurons, so a patient's ability to express emotions and attitudes is less affected than in cases of neurological disorder.

Since the speaker in our case study is an individual with amyotrophic lateral sclerosis (ALS), a subtype of MND, we provide some background information on MND. According to the Motor Neurone Disease Association webpage (Motor Neurone Disease Association, n.d.), MND is an embracing term used to cover a number of progressive illnesses that affect motor neurons in the brain and spinal cord. Subtypes include ALS, progressive muscular atrophy (PMA), progressive bulbar palsy (PBP), and progressive lateral sclerosis (PLS). The dominant symptom is the weakening and wasting of muscles, resulting in impaired physical abilities, while the intellect and emotions remain unimpaired. Respiratory muscles also weaken eventually, and many individuals need to undergo tracheotomies that in most cases result in losing the ability to speak. MND can affect adults at any age but most people who contract the disease are in the 40–70 age range. Generally, men are affected slightly more often than women. The precise figures for the incidence and prevalence of MND are still uncertain. The incidence (i.e., number of people who develop MND in any one year) is approximately 2 per 100,000, while the prevalence (i.e., number of people who actually have MND at any one time) is thought to be approximately 7 per 100,000.

Patients and their families describe how difficult and painful it is not to be able to communicate. The fear of losing the ability to communicate is so great that in some cases, patients refuse a tracheotomy altogether, even if the consequences entail death (Toyoura, 1996). Although we recognize that nothing completely fulfills the desire to express oneself in the way in which an individual has grown accustomed, Chatako-AID at least offers the ability to phonate with that individual's own natural speech quality.

3. The Speech Synthesis System Used in This Study

ATR CHATR is a corpus-based natural speech resequencing synthesis system that incorporates read speech as a source database for concatenative unit selection (Campbell and Black, 1997). CHATR maintains the naturalness of the original voice as it reduces the amount of subsequent signal processing by taking advantage of the natural phonetic and prosodic variation of source units. The system runs on various operating systems such as UNIX, Linux, and MS Windows. Among them we used CHATR98, which runs on MS Windows platforms. In our experiments, we did not apply any signal processing to the output speech waveform. The system is built as a one-time off-line process for source database creation, and runs in real-time for the online TTS synthesis process.

3.1. Source Database Creation in CHATR

The first step in source database creation is to record an individual's speech and label it phonemically to create a speech corpus. Any individual who can read, from children to the elderly, can serve as speech donors. The next step is to store the resulting speech in digital form on a PC. At this stage, processing takes place to eliminate disfluencies and redundancies within the speech, to connect the speech waveforms, and to store them externally as a source database, which is usually separate from the synthesizer. An inventory file for access to the source database is created in a one-time off-line process consisting of the following three steps:

- convert an orthographic transcription of the corpus texts to an equivalent phonemic representation;
- align the individual phonemes to the waveform to provide a start time for each unit so that the prosodic features can be measured;
- produce prosodic and contextual feature vectors for each unit.

Information such as phoneme label, start time information, duration, fundamental frequency (indicated as F0, commonly referred to as "pitch"), probability of voicing, and RMS energy are included as features. From the start time information, phoneme labels of neighboring units can be identified in the inventory (Fig. 1 is an example inventory file).

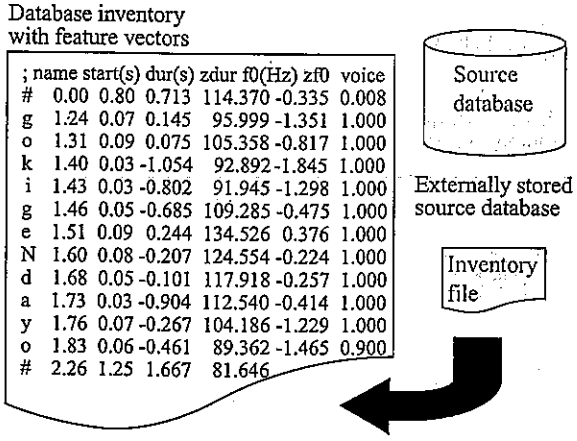

Figure 1. Example of inventory file. Start(s): starting time, dur: duration, zdur: z-score of duration, f0: fundamental frequency (F0), zf0: z-score of F0, voice: probability of voicing.

The quality of the synthesized speech depends on the richness of phonetic and prosodic variation in the speech corpus as well as on the recording quality. The ideal size of a corpus has not yet been determined but, in past trials, the read speech of a phonetically balanced ATR sentence set comprised of 503 isolated sentences (Abe et al., 1990) successfully served as a source database (Black and Campbell, 1995). ATR later prepared a richer sentence set (hereafter, ATR 525) with 22 sentences added to the original ATR corpus, which can be read in about an hour or two by a healthy adult. The phonetic balance of a corpus is measured by counting how many biphone (i.e., two phone sequences such as /ak/ or /da/) combinations appeared in that corpus. The profile of ATR 525 is thus 525 sentences, 16,612 morae (a unit in Japanese corresponding to a single vowel or a consonant-vowel combination), 31,053 phonemes, and 403 biphone combinations.

3.2. *The CHATR TTS Synthesis Process*

After CHATR receives a text as input, it takes three steps in the real-time online TTS synthesis process: (1) text analysis (text-to-phoneme conversion, accent tagging, and break indexing); (2) prosodic prediction; and (3) unit selection. Figure 2 illustrates the CHATR TTS synthesis process. This section describes the prosody prediction and unit selection processes.

3.2.1. Prosody Prediction in CHATR. The standard CHATR predicts F0 and duration by making use of linear regression (LR) statistical models that were trained

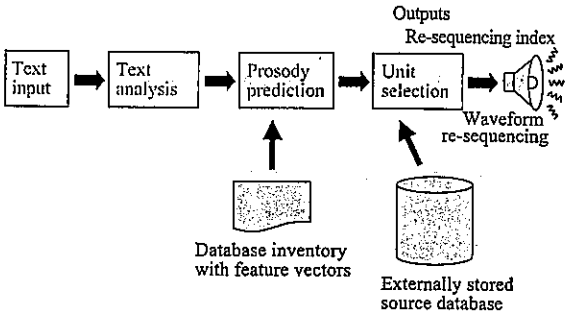

Figure 2. CHATR TTS synthesis process.

on the reading of the original ATR text corpus and on other materials produced by a male professional announcer (hereafter, the model speaker) of standard Japanese in his normal reading style. One LR model per unit was trained for duration prediction and three LR models per mora were trained for F0 prediction; at the start, mid-point, and end. The models predict duration and F0 values from J-ToBI labels—a labeling scheme based on the Japanese Tones and Break Index prosodic labeling system proposed by Venditti (1995) as an extension of ToBI, which was originally designed for English—given to the input text (Black and Hunt, 1996; Campbell, 1996). The predicted F0 and duration values are then mapped to equivalent features in the source database by means of normalized values (z-scores), from their means and standard deviations (SD). Normalized values are then converted to absolute values, referring to a table of means and standard deviations for the source speaker, which are stored as a part of the source database. In other words, the predicted pattern serves as a guideline and CHATR automatically finds the optimal sequence of speech units for each speaker by selecting the concatenation units closest to the predicted patterns from the available units in the source database. In this way, standard patterns predicted with the LR models can be mapped to any source database by the transformation, enabling speaker-specific prosodic phrasing to be produced.

3.2.2. The Unit Selection Algorithm in CHATR. CHATR first compiles a list of candidate units per phone from the source database inventory, looking up feature vectors and neighboring unit information (Black and Campbell, 1995). It then selects the optimal unit to create a target utterance by maximizing continuity and minimizing the distance from phonetic and prosodic targets. Two cost functions are used in this process. One is a target cost, the degree to which units

match the target specification, and the other is a concatenation cost, the degree to which two adjacent units can be imperceptibly concatenated. As the last stage in the process, CHATR produces an index file of the optimum units in sequence and accesses an externally stored source database for waveform re-sequencing according to the time index information in that file.

4. Creation of the Speech Corpora

The objective of the speech corpus design in this study is to reproduce the characteristics of the speech of one particular individual. Here, characteristics refer to this person's natural intonation, articulation, and voice quality. We created three kinds of speech corpora in Japanese for 'neutral' speech (i.e. in a normal speaking style—hereafter, main source database). In addition, we created speech corpora marked for three kinds of emotion.

4.1. The Speaker for the Case Study

We introduce the speaker's name and personal information with his permission and at his request. Mr. Shinichi Yamaguchi, who is 64, resides in Fukuoka, Japan, and was diagnosed with amyotrophic lateral sclerosis seven years ago. At the time of diagnosis, he already had difficulty with spontaneous respiration. Since then, he has been wearing a nasal pressure support ventilator 24 hours a day. Formerly, he was an electrical engineer and taught computer science at the college level. Since being diagnosed with ALS, the speaker has been active in giving talks to public audiences on the effectiveness of computers for people with disabilities. He is aware of the possibility of losing his voice in the future, and he thinks that the speech synthesized by current commercial systems sounds less natural than the human voice. As a result, he hopes to use more human-sounding, expressive speech generated from his own voice (Yamaguchi, 2000). He has shown a keen interest in our research and collaborated with us as both a speaker and a user.

4.2. Text Corpora for Main Source Database

Maintaining phonetic balance is important to assure that the system is able to synthesize all phones in various phonetic contexts. There has been, however, a good deal of concern that phonetically balanced sentences are both difficult to read and remote from the style of utterances used in daily conversation. Therefore, we decided to use reading materials familiar to (and written by) the speaker, assuming that the reading of such materials would exemplify the speaker's own natural intonation, articulation and voice quality. In Iida et al. (2003), we confirmed that the phonetic balance of the corpora created from familiar texts closely matched the phonetic profile of the ATR 525 sentences described in Section 3.1.

The following section describes the three text corpora that we used for the main source database in this study.

4.2.1. Balanced Text Corpus. A phonetically balanced sentence set was used to ensure that the system would be able to synthesize all phoneme combinations. We extracted 129 sentences from the ATR 525 based on the criterion of at least one appearance for all the necessary biphone combinations for Japanese composition. The purpose of making the reduced subset was to lighten the speaker's recording load. The number of biphone combination was 465.

4.2.2. Speaker Text Corpus. A manuscript written (and regularly read) by the speaker (Yamaguchi, 2000) was used as a source of natural intonation, articulation, and voice quality for running speech. We expected more natural phonation and prosodic patterns to be found in readings of the familiar manuscript. This text set contained 348 sentences with 385 biphone combinations.

4.2.3. Daily Text Corpus. One of the requirements for a VOCA is the accurate production of frequently used words and phrases. Therefore, we used a list of words and short sentences that Mr. Yamaguchi himself prepared for his own use. The idea was that such material would include units in a continuum from the source database, and would therefore reduce the discontinuity of concatenation when he mixed the words and sentences from such materials. The words and sentences includes requests to caretakers, conversations with caretakers and friends, conversations on the phone, and words essential to daily conversation (i.e., parts of the body, symptoms, directions, etc.). This set contained 91 short sentences and 495 words.

4.3. Emotionally-Colored Texts Corpora

Joy and sadness were the themes of monologues taken from autobiographies of the disabled, and angry

monologues were chosen from the speaker's writings. The first author modified published texts (with permission from the original authors) to elicit particular emotions, and added 50 sentences from the ATR 525 original to ensure that at least one appearance for all the necessary biphone combinations would appear in each text corpus. There were 138 sentences used for recordings of anger, 185 for joy, and 141 for sadness.

4.4. Recording of Speech Corpora

We created a speech corpus that corresponds to the text corpus by having the speaker read it aloud. Recording was scheduled over two days and took place in a barrier-free sound-proof room. Mr. Yamaguchi was accompanied by his wife and a volunteer recording staff.

The speaker's nasal pressure support ventilator is designed to give alternating high and low pressure. When high, the ventilator pressures the user to aspiration and when low, expiration. The ventilator emits a motor noise while producing high pressure. To reduce the noise, we asked the speaker not to speak while in aspiration and to speak only while the ventilator was in low pressure. We also covered the ventilator with a blanket to reduce noise. We paid particular attention to the speaker's health conditions, allowing plentiful rest when needed. During the recording of emotional speech, the experimenter (the first author) periodically introduced conversations on topics intended to help induce the target emotion. For details of emotion elicitation during the recording see Iida et al. (2003). Because we gave priority to recording materials for the main source database, we had enough time to record only one-third of the target sentences for each emotion. Thus, the emotional speech synthesis is not yet ready for synthesizing fully emotional texts in the current prototype version.

To summarize, we produced the following corpora:

For main database

- Balanced text corpus: Phonetically balanced sentences (129 sentences)
- Speaker text corpus: Familiar texts for the speaker (348 sentences)
- Daily text corpus: Words and sentences used daily (91 short sentences, 459 words)

For Emotional speech

- Anger text corpus (138 sentences)
- Joy text corpus (185 sentences)
- Sadness text corpus (141 sentences)

5. Evaluation of Speech Synthesized Using the Corpora

We created four kinds of source database with one or more corpora. Using each source database, we synthesized speech from six test sentences with CHATR. We observed selected units, measured objective distances, and performed perceptual experiments with 20 Japanese informants in order to decide which source database to use for the VOCA implementation. We then evaluated our method's feasibility for practical use by performing perceptual experiments with a commercial equivalent with twenty different informants.

5.1. Source Databases for Evaluation

Each source database was constructed with the following four corpora:

- DB1: Balanced corpus
- DB2: Speaker corpus
- DB3: Combination of Balanced and Speaker corpora
- DB4: Combination of Balanced, Speaker and Daily corpora

5.2. Sentences Used for Evaluation

Six test sentences were prepared and used in all evaluations. To maintain impartiality, we did not use words and sentences from the daily corpus to compose the test sentences. We subcategorized sentences into three types: (1) two sentences containing numbers since numbers are important and need to be perceived correctly; (2) two sentences containing uncommon noun collocations with the assumption that pairs of uncommon collocations would discourage guessing; and (3) two sentences containing emotional expressions since that are considered to be important for daily conversations. Table 1 shows the test sentences with English translations, and the ratio of units selected from each source database. The test sentences were synthesized with CHATR using each source database. All speech samples were stored as 16-bit Microsoft wav format files using a 16 kHz sampling rate.

5.3. Analysis of Synthesized Speech Created with the Four Source Databases

DB1 and DB2 each consisted of a single speech corpus, so where units were selected from was obvious. When

Table 1. Sentences used for evaluation. "U" denotes a sentence with uncommon noun collocations, "N", a sentence with numbers, and "E", a sentence with emotional expressions.

Sentence Type	Sentence Group	Num. of syllables	Sentence
U	X	4	Akira-kunwa/yamazakurato/kakinabeto/ iimashita. (Akira said a mountain cherry tree and an oyster pot).
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).
E	X	5	Wa-i,/yatto/kurundane./Ganbattekite/ yokatta. (Wow, so he is finally comming! I'm glad I've been trying so hard.)
U	Y	4	Yamamoto-sanwa/Kaichudentoto/ kureyonto/iimashita. (Ms. Yamamoto said a torch and a crayon.)
N	Y	5	8gatsu15nichino/nichiyoubikara/ 8gatsu/16nichino/getsuyoubi./ ippakufutsukadesu. (From Sunday, August, 15 from Monday, August, 16, a one night two days trip.)
E	Y	5	Ah,/tsukareta./Hajimetenanode/ totemo/fuandesu. (Oh, I am tired. I am very worried since it is my first time.)

sentences were synthesized with DB3, the ratio of units selected from the Balanced corpus to those selected from the Speaker corpus varied from 24.5 to 75.5% for the six speech samples. The longest sequence taken from the Speaker corpus was 4 units and the average 1.8 units. Selections from the Balanced corpus were equivalent. When synthesized with DB4, the ratio of units selected from the Balanced corpus, the Speaker corpus, and Daily corpus was 18.1% to 74.1 to 7.8%. The longest sequence of units taken from the Balanced corpus was 7 and the average 1.9, while equivalent averages for the Speaker and Daily corpora were 5 and 1.9, and 4 and 1.6, respectively. Results are shown in "open text" row in Table 2. An example of unit selection using DB3 and DB4 is shown in Table 3. The sentence used was "Hajimetenanode, totemo fuan desu (I am very uneasy, since it is my first time)".

For evaluation, we measured F0 distance by referring to the *distance of slope* ($\Delta^2 F_{rms}$) in Marumoto and Ding (1998) with the following equation:

$$\Delta^2 F_{rms} = \sqrt{\frac{\sum_{i=1}^{n} (\Delta F_t(i) - \Delta F_s(i))^2}{n-1}} \quad (1)$$

$F_t(i)$ is the F0 value of the i-th target unit, $F_s(i)$ is the F0 value of a selected unit equivalent, $\Delta F_t(i)$ is the distance between the F0 of the i-th target unit and its immediately previous target unit, and $\Delta F_s(i)$ reflects the equivalent distance between the F0 of the selected i-th unit and its immediately previous selected unit. We measured the root mean square (rms) of the difference between the two values. The result indicates an overall approximation of F0 slope between target and selected units for the whole sentence. We obtained results as shown in Table 4 with DB1: 23.0 Hz, DB2: 20.5 Hz, DB3: 23.3 Hz, and DB4: 23.9 Hz.

For duration distance, we measured *rms error* (ΔD_{rms}) using the following equation:

$$\Delta D_{rms} = \sqrt{\frac{\sum_{i=0}^{n} (D_t(i) - D_s(i))^2}{n}} \quad (2)$$

where $D_t(i)$ denotes the duration of i-th target unit and $D_s(i)$ denotes the duration of i-th selected unit.

Table 2. Ratio, longest sequence, average of units selected from each corpus. Balance: Balance corpus, Speaker: Speaker corpus, Daily: Daily corpus.

		DB3		DB4		
		Balance	Speaker	Balance	Speaker	Daily
Open text (6 sentences in Table 1)	Ratio (%)	24.5	75.5	18.1	74.1	7.8
	Longest (units)	4	4	7	5	4
	Average (units)	1.8	1.8	1.9	1.9	1.6
Task-oriented Text	Ratio (%)	21.6	78.4	8.1	45.6	46.3
	Longest (units)	6	2.2	3	7	1.2
	Average (units)	3	1.3	1.7	2.5	3.5

Table 3. Example of selected units when synthesized using Database 3 and 4 for the sentence, "Hajimetenanode totemo fuan-desu (I am very unsasy, since it is my first time)." "#" a pause between phrases, "U" an unvoiced "U", and "N" a mora (syllabic) nasal.

DB3		DB4	
Speaker corpus345	h	Balanced corpus047	h
Speaker corpus037	a j i m e	Speaker corpus037	a j i m e
Speaker corpus009	t	Speaker corpus228	t
Speaker corpus079	e	Daily corpus058	e n
Balanced corpus010	n	Speaker corpus055	a n o d
Speaker corpus055	a n o d	Speaker corpus238	e
Speaker corpus188	e #	Speaker corpus205	# t
Speaker corpus147	t o t	Speaker corpus298	o t
Speaker corpus345	e m	Speaker corpus165	e m o
Balanced corpus027	o #	Speaker corpus347	#
Balanced corpus026	f	Daily corpus004	f
Speaker corpus345	u a	Speaker corpus345	u a
Speaker corpus259	N d	Speaker corpus259	N d
Speaker corpus145	e s U	Speaker corpus047	e s
		Speaker corpus044	U

Table 4. Distance, max of difference, intelligibility and MOS for each database.

	DB1	DB2	DB3	DB4
Distance of F0 slope (Hz)	23.0	20.5	23.3	23.9
RMS error for duration (ms)	24.8	19.3	21.3	27.4
Intelligibility (%)	66.9 ± 31.6	65.4 ± 33.7	92.7 ± 13.6	87.4 ± 20.7
MOS (Highest: 5, Lowest: 1)	2.7 ± 1.2	2.4 ± 0.9	3.2 ± 0.9	3.2 ± 0.8

The result of this measure indicates the overall approximation of the duration between target and selected units for the whole sentence. The results were DB1: 24.8 ms, DB2: 19.3 ms, DB3: 21.3 ms, and DB4: 27.4 ms (see also Table 4).

For both measures, the smaller values indicate that the two parties (i.e., target and selected units for the whole sentence) are in approximation. Since the values do not approximate zero, the results suggest that there are some distortions in speech synthesized with any source database. Among four source databases, the values of DB2 were the smallest for both measures. The primary cause of this result may be attributed to the fact that the DB2 was comprised of only the Speaker corpus that was read naturally by the speaker as a source for running speech, but the point needs further observation. Results of distance measures are summarized in Table 4 along with the results of perceptual experiments described in the next section.

5.4. *Perceptual Evaluation of Synthesized Speech Produced with the Four Source Databases*

To evaluate the intelligibility of speech synthesized from the four source databases, we conducted perceptual experiments with 20 Japanese student and colleague informants. Sentences in Table 1 were synthesized with four kinds of source database, which resulted in 24 speech samples. Each of the 20 informants listened to 4 speech samples. For all informants, each sentence was synthesized with a different source database and each sentence-database combination was presented only once to each informant. Thus, five responses were obtained for each sentence/database combination.

Informants were asked to type all words exactly as they heard them. The intelligibility of each sentence was measured by totaling the correct number of *bunsetsu*. A bunsetsu is a syntactic unit in Japanese consisting of one content word, e.g., a noun or verb that imparts meaning by itself, followed by one or more functional morphemes, such as auxiliary verbs or prepositional phrases. The sentences used for the experiment contained between five and seven bunsetsu units.

The overall mean intelligibility score (as a percentage) and its standard deviation (SD) for each source database was as shown in Table 4 with DB1: 66.9 ± 31.6%, DB2: 65.4 ± 33.7%, DB3: 92.7 ± 13.6%, and DB4: 87.4 ± 20.7% (written as mean ± SD).

We performed means comparisons for all pairs using a Tukey-Kramer HSD (Honestly Significant Difference) of $p < 0.05$ (Sall and Lehman, 1996). As Fig. 3 shows, there was no significant difference either between DB1 and DB2, or DB3 and DB4, although there was a significant difference between the two groups (i.e. between the following combination: DB1/DB3, DB1/DB4, DB2/DB3, and DB2/DB4). This result indicates that source databases comprised of 2 or 3 corpora were more intelligible than those comprised of only 1 corpus.

In addition, we asked the twenty informants to evaluate speech synthesized using all 4 source databases to give their subjective impressions of each sample using

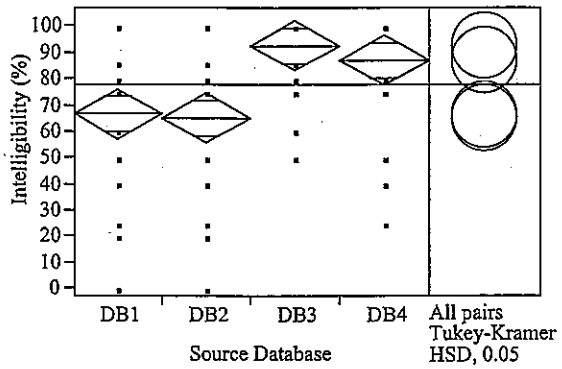

Figure 3. Mean diamonds and comparison circles showing the result of intelligibility test. The horizontal line across the middle of both graphs shows the overall mean of all the observations. Centerlines of the means diamonds are the group means, lines near the bottom and the top are 95% overlap marks, and the bottom and the top of the diamonds form the 95% confidence intervals for the means. The center of each comparison circle is aligned with its group mean and the radius of a circle is the 95% confidence interval for its group mean.

a 5-point scale (5 = excellent, 1 = poor). The objective of this evaluation was to ascertain the informants' likes and dislikes regarding the quality of the synthesized speech. The mean opinion scores (MOS) and SDs for the speech synthesized with each source database were DB1: 2.7 ± 1.2, DB2: 2.4 ± 0.9, DB3: 3.2 ± 0.9 and DB4: 3.2 ± 0.8 (also see Table 4).

We performed the same statistical analysis as we applied to the intelligibility test. As Fig. 4 shows, there was no significant difference among DB1, DB3, and DB4, or between DB1 and DB2. Significant differences were seen, however, between DB2 and DB3, and between DB2 and DB4.

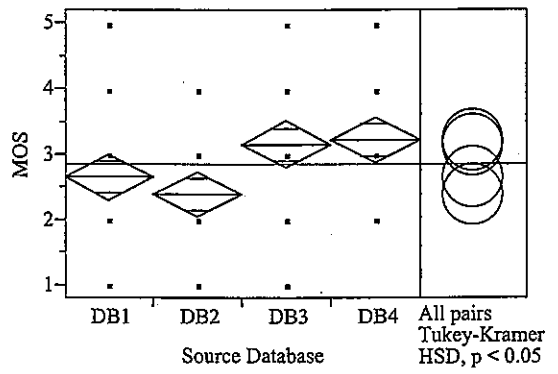

Figure 4. Mean diamonds and comparison circles showing the result of subjective impression test. For graphical explanation, see caption of Fig. 3.

Results of the perceptual evaluations suggest that DB3 and DB4 are better choices for the VOCA implementation, although distance measures suggest that there are least distortions in speech synthesized by DB2. Consequently, no correlation was observed between results of the objective distance measures and the perceptual evaluation. Further investigation using more sophisticated acoustic transformations (e.g., Chen and Campbell, 1999) will be necessary, but for this study, we gave priority to the results of the perceptual experiments.

5.5. *Source Database for VOCA Implementation*

One of the requirements for a VOCA is the accurate production of frequently used words and phrases. We selected DB4 (over DB3) as a source database for Mr. Yamaguchi's VOCA because this source database included the Daily text corpus that consisted of a list of words and short sentences he himself prepared for his own use. CHATR has a bias to select contiguous units, even with the current phone-based unit selection algorithm. We considered the tendency an advantage, since the addition of DB4 would enable CHATR to reproduce the natural quality of the words and sentences in DB4. As shown in Table 2 (for the task-oriented text), we confirmed this tendency by synthesizing 25 short sentences included in the daily corpus. When synthesized with DB4, the ratio of units selected from the Daily corpus, Speaker corpus, and Balanced corpus was 46.3 to 45.6 to 8.1% respectively. The longest sequence of units taken from the Daily corpus was 12 and the average 3.5, while equivalent figures for the Speaker corpus were 7 for the longest and 2.5 for the average. For the Balanced corpus, the longest sequence was 3 units and the average 1.7. When synthesized with DB3, the ratio of units selected from the Speaker corpus to those selected from the Balanced corpus was 78.4 to 21.6%. The longest sequence of units taken was 6 from the Speaker corpus and the average 2.2. Equivalent figures for the Balanced corpus were 3 for the longest sequence and 1.3 for the average.

The advantage of selecting contiguous units is that it successfully carries the pitch contour of natural speech onto the synthesized speech. For example, Fig. 5 shows the pitch contours of the phrase, "Kyuuin shite kudasai (please remove the saliva from my mouth)" for the natural speech and the synthesized speech using DB4 and DB3. As observed, the pitch contour for speech synthesized with DB4 resembles that of natural speech more

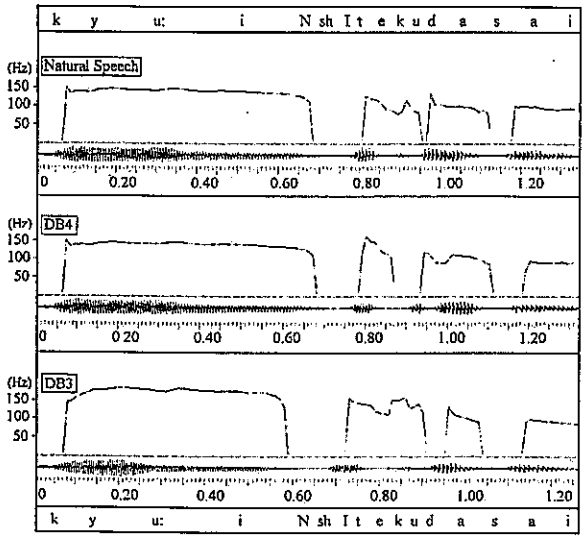

Figure 5. Pitch contours of "Kyuuin shite kudasai (Please remove the saliva from my mouth)." From the upper layer, natural speech, speech synthesized with DB4, and speech synthesized with DB3. F0 values are extracted and plotted by WaveSurfer 1.4.7 (KTH Department of Speech, Music and Hearing, 2003). This figure reproduced pitch contours, waveforms and time axes displayed by WaveSurfer.

closely than the pitch contour synthesized with DB3. This is because, as shown in Table 5, units for "kyuuin" were selected as a whole from the Daily corpus when synthesized with DB4. The capability of reproducing the natural intonation is important, especially when it comes to crucial terms such as "kyuuin."

5.6. *Perceptual Evaluation on a Practical Level*

To evaluate the feasibility for practical use when our synthesis method is implemented in a VOCA, we performed perceptual experiments to compare our software, using DB4, with a commercial equivalent: Ricoh's Yuubenka ver.2, released in June, 1997 (Ricoh Co. Ltd., n.d.) with 20 informants. None of the informants who participated in this follow-up experiment participated in the source database comparison tests described in Section 5.4. Informants evaluated intelligibility and provided their subjective impressions following the same procedures as for tests in Section 5.4. Ricoh's Yuubenka ver. 2 is a CV + VC concatenative synthesis system (where C represents a consonant and V a vowel) that applies signal processing after concatenation. We selected this synthesis system, which operates on MS Windows 95 and higher, since its LSI version (Ricoh RL5S850, released in April, 1998) was implemented in Namco's Talking Aid (Namco Co. Ltd., n.d.), a VOCA released in May 1999 and used widely in Japan. We synthesized sentences in Table 1 with the Yuubenka ver. 2 and stored the speech in the same file format as speech previously synthesized by our method using DB4 for the tests described in Section 5.4: i.e., 16-bit wav format files at 16 kHz sampling (cf. Section 5.2). Based on our judgment, Yuubenka 2's intonation, speed, and pitch were set to the most natural-sounding level, and the volume was adjusted to the same level for all sentences.

To avoid the semantic influence of text content on the speech samples, we separated the 20 informants into two groups of 10. The six sentences were also separated into two groups (X, Y) of three sentences each. The first informant group listened to speech samples of group X synthesized by our software and group Y synthesized by the commercial product. The second informant group listened to the reverse combination of speech samples (i.e., group Y synthesized by our software and group X synthesized by Yuubenka). Thus, 10 responses were obtained for each speech stimulus. The overall mean intelligibility score and SD were $92.0 \pm 15.7\%$ for our method and $92.9 \pm 13.1\%$ for the commercial system. With a t-test of $p < 0.05$, no significant difference was recognized. However, for subjective impressions of informants, the MOS and SD were 3.4 ± 0.9 for our system and 2.9 ± 1.0 for Yuubenka. Mr. Yamaguchi and his wife also participated in the experiment. The speaker's score was 91.7% for our method and 93.3% for the commercial system, while his wife's score was 88.6% for the former and 93.3% for the latter. On a subjective impression test both gave our method a 3.7 rating. The speaker gave the commercial system a 2.3 and his wife a 2.7. We

Table 5. Selected units in sequence when synthesized with Database 3 and 4 for "Kyuuin shite kudasai (please remove the saliva from my mouth)," an important word for target users selected from DB4. "I" denotes an unvoiced "i", "u:" a long "u" vowels, and "N", for mora (syllabic) nasal.

DB3		DB4	
Balanced corpus037	k	Daily corpus011	ky u: i N sh
Speaker corpus028	y u:	Speaker corpus125	I t
Balanced corpus027	i N sh	Speaker corpus055	e k
Speaker corpus065	I t e k	Daily corpus019	u d a s
Speaker corpus039	u d a s a i	Daily corpus009	a i

infer that these results confirm our synthesis method to be suitable for practical use, comparable in intelligibility and superior in preference to the commercial system.

6. Development of Chatako-AID

The primary advantage of Chatako-AID using CHATR and DB4 is, as already described, that it synthesizes text to speech using the speaker's own voice. In addition, Chatako-AID utilizes an acceleration function that enables the speaker to select a word or a short sentence from a pre-stored list to produce the corresponding speech segment or to paste the text in a text input window for CHATR to synthesize. Chatako-Aid can also used by other target individuals (whom we described in Section 2) by loading source databases created with their speech. Furthermore, Chatako-AID supports multilingual use if there is a source database available in the target language. Currently, Chatako-AID's graphical user interface (GUI) is implemented for both Japanese and English.

6.1. *System Configuration*

At present, Chatako-AID runs only on Windows using CHATR98 but could be easily ported to other platforms. Figure 6 shows the configuration of Chatako-AID. The GUI for Chatako-AID is implemented in Tck/TK 8.3 (a platform-independent script programming language) with a text window, speaking style selection menu, command buttons, and selectable list of pre-stored words and sentences.

6.2. *Main Window*

The top half of Fig. 7 shows the English-mode main window at the initial stage. On the top left is a pull-down menu for changing speaker's voice (changing the current source database to the one made with different speaker's voice). Below the menu, a text window is located on the left, with command buttons on the right. The following eight commands are implemented: *synthesize*, *paste text*, *open text file*, *save wave*, *save text*, *clear screen*, *stop synthesis* and *exit*. The language on the screen can be changed automatically according to the selected source database.

Located above the command buttons are icons for speaking style selection. We positioned *neutral* at the leftmost and then *joy*, *anger*, and *sadness* from left to right. The default is set to *neutral* (the main source database, DB4). The user selects an appropriate speaking style before typing the text, which is then distinguished by different font colors (*joy* by green, *anger* red, and *sadness* blue). The user can change speaking styles by highlighting typed text and then re-selecting the desired speaking style, which leads to the switching of current database to the other database with the desired speaking style.

6.3. *List Window*

Below the text window and command buttons is a selector bar. When clicked, a list of pre-stored words and short sentences appear (as shown in the bottom half of Fig. 7). The list and its format and ordering can be tailored to each user's preference. For the prototype, we prepared a three-layered list. The left column is the

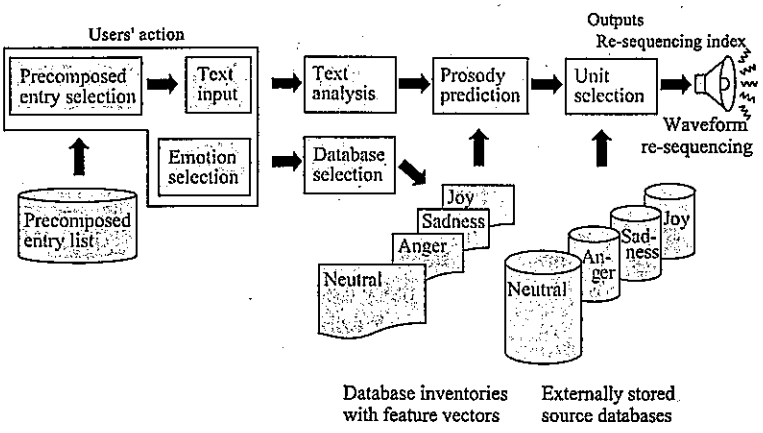

Figure 6. System configuration of Chatako-AID.

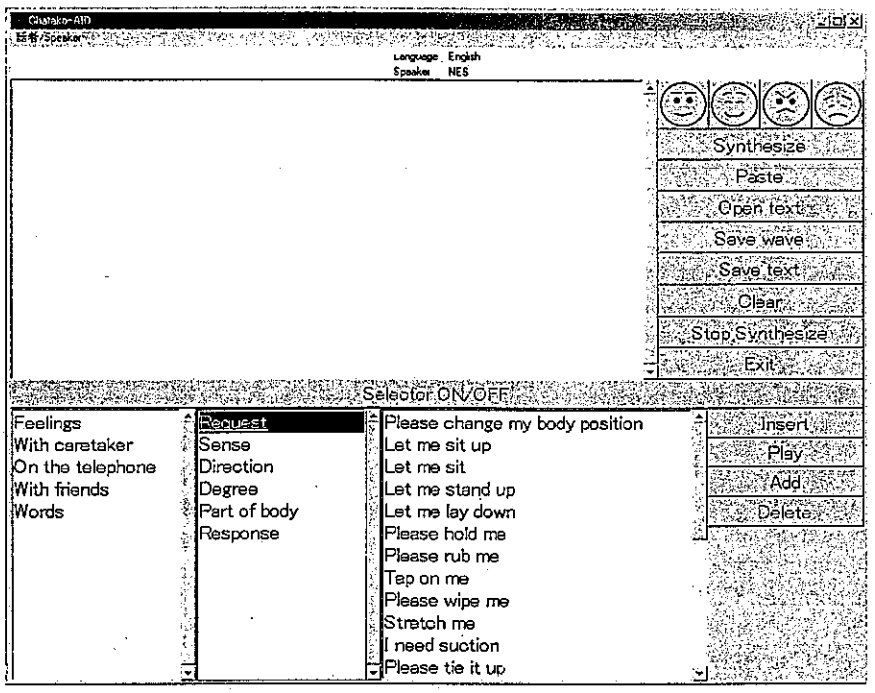

Figure 7. English-mode main window and list window of Chatako-AID.

top layer classified according to addressee. The middle column is the second layer classified according to content (request, greeting, etc.), and the right column is the third layer that lists the words and short sentences to be synthesized.

Command buttons are located on the right. The *play* command synthesizes the selected entry, while the *insert* command copies the entry to the text window for CHATR to synthesize as part of a longer utterance. The *add* command creates a pop-up window for adding words and short sentences to the user-defined list. The command at the bottom, *delete*, removes an entry from the list. The same procedures add and delete category labels in higher layers as well.

7. Future Work

Further research on speech database design, coverage, and balance needs to be carried out to improve the quality of the speech synthesized by this method. It is also important to develop a method for reducing the prosodic and phonetic distortions at unit boundaries before and after selected whole-word segments.

Another area of research that we feel needs to be pursued is the enhancement and selection of pre-stored text utterances, which will allow the system to operate more effectively as a VOCA. We plan to monitor target users, in order to better categorize their utterances into a core and an extended vocabularies. Lastly, we recognize that the input acceleration function should be improved. We also plan to redesign the GUI by incorporating graphic symbols for more intuitive user access.

8. Conclusion

In this paper, we have presented a study of one individual anticipating the loss of phonatory function for whom we created a source database for a corpus-based speech synthesis system (ATR CHATR). Results confirmed that our method can synthesize intelligible speech with natural speech quality. Our speaker has now been using the system for a year to give public presentations and to conduct the various interactions of his daily life, although at the present moment he can still speak (albeit with some difficulty) using his own voice. We feel that our findings present encouraging results for other individuals facing the loss of their voices.

In summary, we created four kinds of source database, combining 1 to 3 speech corpora from an ALS

patient's read speech. Using each source database, we synthesized speech from 6 test sentences with CHATR. We observed selected units, measured objective distances, and performed perceptual experiments with 20 Japanese informants in order to decide which source database to use for the VOCA implementation. Distance measures failed to show clear differences among the source databases. However, perceptual evaluations confirmed that the speech synthesized with source databases composed of more than one corpus (DB3 and DB4) was more intelligible and produced a more favorable impression than speech synthesized with source databases comprising a single corpus (either DB1 or DB2). Further analysis showed that DB4, which contains words and sentences frequently employed by the user, was successful in synthesis approximating the natural speech of the speaker. We performed perceptual experiments comparing our synthesis method with a commercial software program with twenty informants who did not participate in the source database comparison tests. The results showed that speech synthesized by our method was as intelligible as the speech synthesized by the commercial system while receiving more favorable responses with respect to vocal quality. Building on our results, we developed a VOCA incorporating CHATR and the selected source database as an AAC device for the speaker with an input acceleration function. In addition, we designed our system for loading emotional speech separately from the neutral speech database. Future directions of research include improving speech quality and enhancing the input/output acceleration function.

Acknowledgments

The authors would like to express their sincere appreciation to Mr. Shinichi Yamaguchi of Fukuoka, Japan for his participation in the research. We are grateful for financial assistance from the Japan Science and Technology Agency via the CREST (Core Research for Evolutional Science and Technology) scheme for Advanced Media Technology. We are grateful to Prof. Kimitoshi Fukudome of the Kyushuu Institute of Design for providing such excellent recording facilities. Further appreciation goes to Professors Satoshi Imaizumi and Keikichi Hirose of the University of Tokyo for recording advice. We also thank ATR for providing CHATR98 for the speaker and Mr. Ken Shimomura of NTT AT for the support on the text corpora design. We are also grateful to the students and colleagues who participated in our perceptual experiments. Lastly, we thank Prof. Michiaki Yasumura and Dr. Fumito Higuchi of Keio University for their valuable advice and support on evaluation procedures.

References

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

Beukelman, D.R., Yorkston, K.M., Poblete, M., and Naranjo, C. (1984). Frequency of word usage in communication samples produced by adult communication aid users. *Journal of Speech & Hearing Disorders, 49*:360–367.

Black A. and Campbell, N. (1995). Optimising selection of units from speech databases for concatenative synthesis. *Proceedings of Eurospeech 95*, Madrid, Spain, pp. 581–584.

Black A. and Hunt. A. (1996). Generating F0 contours from ToBI labels using linear regression. *Proceedings of ICSLP96,* Philadelphia, PA, vol. 3, pp. 1385–1388.

Cambridge Adaptive Communication (2002). *Cambridge Homepage.* Retrieved May 20, 2003 from http://www.possum.co.uk/Cambridge/Index.htm

Campbell, W.N. (1996). Autolabelling Japanese TOBI. *Proceedings of ICSLP96,* Philadelphia, PA, vol. 4, pp. 2399–2402.

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

Chen, J. and Campbell, N. (1999). Objective distance measures for assessing concatenative speech synthesis, *Proceedings of Eurospeech99*, Budapest, Hungary, pp. 611–614.

Conroy, D., Vitale, T., and Klatt, D.H. (1986). *DECtalk DTC03 Text-to-Speech System Owner's Manual,* EK-DTC03-OM-001, Nashua, NH: Educational Services of Digital Equipment Corporation.

Hallahan, W.I. (1996). DECtalk Software: Text-to-Speech Technology and Implementation, Retrieved May 20, 2003 from http://research.compaq.com/wrl/DECarchives/DTJ/DTJK01/

Hitachi Keiyo Engineering and Systems Ltd. (n.d). *Hitachi Keiii Sisutemuzu, "Den no Shin"* [Hitachi Keiyo Engineering and Systems Ltd. "Den no Shin"]. Retrieved May 20, 2003, from http://www.hke.co.jp/products/dennosin/denindex.htm (in Japanese)

Iida, A., Higuchi, F., Campbell, N., and Yasumura, M (2003). A corpus-based speech synthesis system with emotion, *Speech Communication, 40*:161–187.

KTH Department of Speech, Music and Hearing (2003). WaveSurfer. Retrieved May 20, 2003 from http://www.speech.kth.se/wavesurfer/

Marumoto, T. and Ding, W. (1998). ATR Technical Report TR-IT-0276 Improving Prosody of CHATR Output Speech Based on Partial PSOLA and a MOS Decision Tree, ATR Interpreting Telephony Research Lab. (In Japanese).

Motor Neurone Disease Association (n.d.) *What is MND*? Retrieved May 20, 2003 from http://www.mndassociation.org/full-site/what/index.htm

Namco Co., Ltd. (n.d.). *Welfare.* Retrieved May 20, 2003 from http://www.namco.co.jp/welfare/disabled/index.html (in Japanese).

National Institute of Neurological Disorders and Stroke (2001). *NINDS Muscular Dystrophy (MD) Information Page; NINDS Motor Neuron Diseases Information Page.* Retrieved May 20, 2003 from http://www.ninds.nih.gov/health_and_medical/disorders/md.htm; http://www.ninds.nih.gov/health_and_medical/disorders/motor_neuron_diseases.htm

Ricoh Co., Ltd. (n.d.). *News Release.* Retrieved March 10, 2003 from http://www.ricoh.co.jp/release/soft/yuben/ (in Japanese).

Sall, J. and Lehman, A. (1996). *JMP Start Statistics.* SAS Institute, Belmont, CA: Duxbury Press.

Toyoura, Y. (1996). *Inochi no Komyunikeishon [Communication between lives].* Touhou Shuppan, Osaka, Japan (in Japanese).

University of Nebraska-Lincoln, Aphasia Group (n.d.). *Using Augmentative and Alternative Communication with People with Aphasia.* Retrieved May 20, 2003 from http://www.unl.edu/aphasia/AAC.html

Vandelheiden, G.C. and Kelso, D.P. (1987). Comparative analysis of fixed-vocabulary communication acceleration techniques. *Augmentative and Alternative Communication, 3*:196–206.

Venditti, J.J. (1995). Japanese ToBI Labelling Guidelines, Technical Report, Ohio State University, Cleveland, OH.

Yamaguchi, S. (2000). Pasokon wo Tsukaikonasou [Let's Use PC]. Japan Amyotropic Lateral Sclerosis Assosiation, Fukuoka branch Newsletter, No. 6 (in Japanese).

Yorkston, K.M., Dowden, P. A., Honsinger, M.J., Marriner, N., and Smith, K. (1988). A comparison of standard and user vocabulary lists. *Augmentative and Alternative Communication, 4*:189–210.