Application trials for the realization of human-friendly broadcasting

Eiichi Miyasaka PhD,
Atsushi Imai,
Toru Takagi,
Toru Imai PhD,
Akio Ando PhD
NHK Science & Technical Research Laboratories, 1-10-11 Kinuta, Setagaya-ku,
157-8510 Tokyo, Japan
e-mail miyasaka@strl.nhk.or.jp

E. Miyasaka, A. Imai, T. Takagi, A. Ando, Application trials for the realization of human-friendly broadcasting. Gerontechnology, 2001; 1(2): 103 - 110. This paper describes some application systems developed to present human-friendly broadcasting services for the elderly, who at times have difficulty following rapid speech, and for the hearing impaired. These systems include a real-time speech rate converter for the former and a real-time transcription system for the simultaneous subtitling of Japanese broadcast news programs for the latter. In the conversion system the listener can change the speech rate at any time. The pitch of the converted speech is the same as that of the original speech, and its quality suffers little or no impairment. The results of listening tests by elderly observers show that the converted speech, with a slower speech rate than normal, is perceptually preferred. The other system is based on a unique continuous speech recognition system developed by our group. It has been successfully applied to two actual NHK news programs since March 2000.

Key words: speech recognition, subtitling, speech rate, hearing impaired, audibility

In the coming 20 years, one out of every four Japanese persons will be over 65 years old. It will be urgently necessary to provide audio broadcasting services that are comfortable for elderly listeners suffering from hearing impairments due to age. It is known that senescent changes occur with age along all the auditory pathways, including the central auditory nervous system and the auditory periphery[1,2].

Most conventional hearing aids, however, can compensate only for frequency deterioration in the outer and middle ear areas, even though digitalized hearing aids have recently been developed. These are ineffective for elderly persons who are suffering from an inability to follow rapidly uttered speech, because such deterioration may be caused by a disorder in the central nervous system. It is difficult at present to compensate physiologically for these deteriorated functions. At the same time, it is reported that the speech rates of professional announcers have increased from about 350 syllables/min to more than 500 syllables/min over the last two decades[3]. Teranishi[4] performed psychological tests to measure the critical identification rate for young students, obtaining a value of about six syllables/s. According to the results, a speech rate of more than 500 syllables/min (=8.3 syllables/s) will cause hearing problems for elderly listeners.

The new hearing aid system developed by the authors, called the speech rate converter, is designed to compensate for degraded speech intelligibility by slowing the input speech rate, while maintaining the original pitch and timbral personality, with little impairment. It tries to temporally compensate for the reduced functions of the central auditory pathways.

As for those with hearing impairments that prevent them from hearing any speech in a broadcast program, no matter how loudly uttered, captioned broadcasting is effective. Captioned broadcasting of news programs has been widely adopted in the United States of America and in European nations, where operators called 'captioners' can manually enter caption manuscripts in real time. A reason why real-time manual captioning is possible in those nations is that the languages use phonetic characters and the correspondence between sounds and words is clear. The Japanese language, in contrast, uses ideographic characters, so that a certain length of time is required for selection among homonyms through kana-to-kanji conversion. We have developed a news-speech recognition system for creating captioned manuscripts

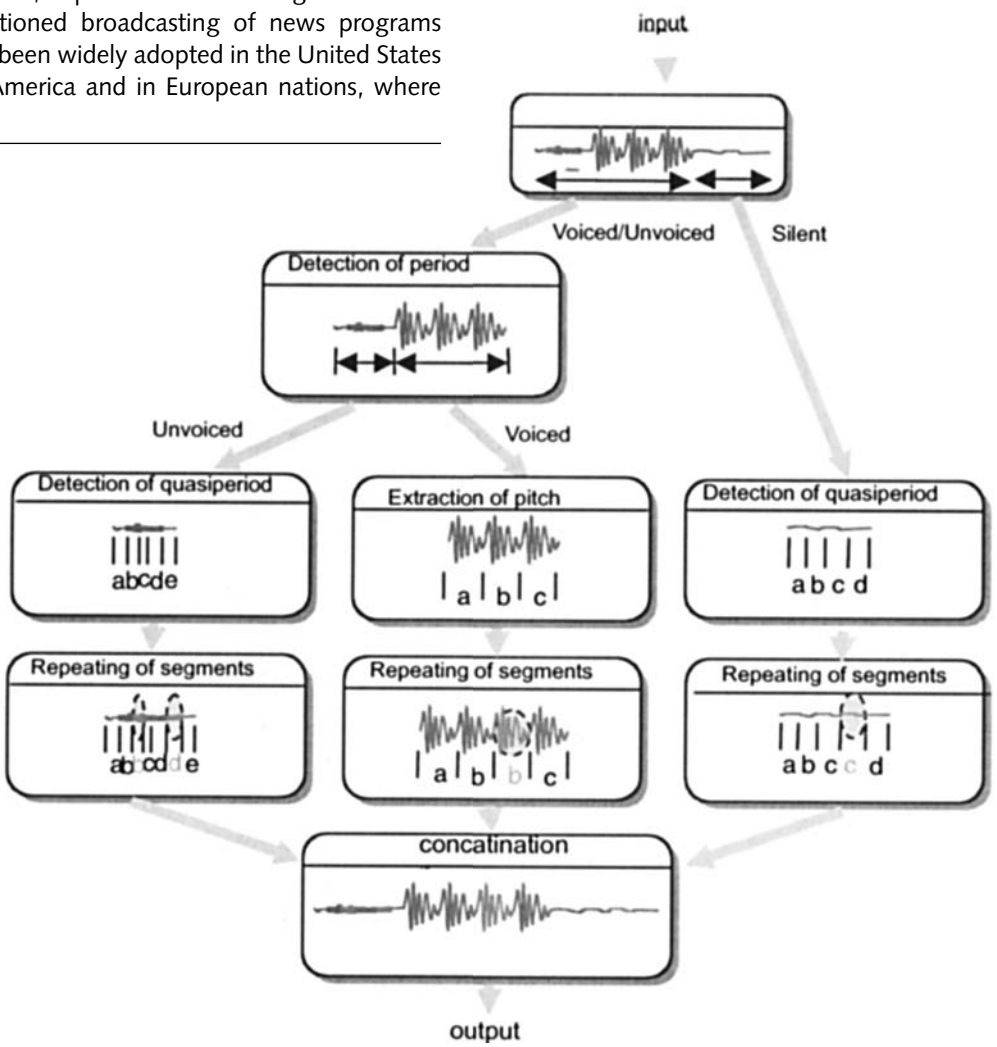

Figure 1. Block diagram of the algorithm of speech rate conversion.

for news programs in real time.
This paper describes both a newly developed real-time speech rate converter and a broadcast news transcription system using the speech recognition system.

REAL-TIME SPEECH RATE CONVERTER

System requirements[5,6]

The main requirements for the new hearing aids for the elderly suffering from rapid speech rate hearing impairment are considered to be the following:

(a) Self-adjustment to the desired speech rate will be indispensable;
(b) The quality of the converted speech should be minimally impaired, because reduced speech articulation will likely cause hearing problems for the elderly. The less impaired the quality, the more effective the system;
(c) Pauses between adjacent breath groups and the voiced portions should be independently handled;
(d) Even in a case where the original speech is accompanied by corresponding video images, like TV, the converted speech would preferably be synchronized with the visual images.

Principle of the system[5]

The speech rate converter presented here satisfies most of the requirements mentioned above. Figure 1 illustrates the block-diagram of the proposed algorithm introduced in this system. Input speech is classified into voiced-, unvoiced-, and silent portions. For the voiced portions, pitch extraction and segmentation are performed as precisely as possible. For the unvoiced portions, as well as the silent portions, extraction of the quasi-period and segmentation are also performed. According to the speech rate specified by a user, enlargement of each portion, whether voiced, unvoiced, or silent, can be performed by a pitch-synchronous insertion of the most suitable number of wave-segments into the original waveforms of the portion. These processed portions are smoothly concatenated in the final stage. The converted speech becomes slower than the original, with no variance in pitch and little impairment in quality. Before enlargement, the processed segments are temporally stored in a large buffer memory. The enlargement process starts as soon as the system receives a trigger, produced 50ms from the moment that a user sets a speech-rate. As a result of this, the speech–rate can be controlled in real-time, even if a user wants to make it faster than the former slowed speech rate.

There are two modes in this system: a linear mode and a nonlinear mode. The former corresponds to the uniform enlargement of the speech at a constant speech rate, while the latter corresponds to a nonlinear enlargement of the speech at a variable speech rate, in order to reduce the discrepancy between the converted speech and the corresponding visual images. The user can select either of the modes. To synchronize the converted speech with the visual images at every onset of a breath group, the speech rate is temporally changed from a slower rate to the normal rate, along a monotone-decreasing curve, similar to the average temporal change in pitch frequency obtained from the speech of veteran announcers. The silent interval between adjacent breath groups can be shortened, if necessary, as long as temporal naturalness is maintained. This function can assist in absorbing the temporal discrepancy during the corresponding breath group.

Subjective evaluations

Subjective evaluations of the system were performed using 356 elderly observers, varying in age from in their 60's to in their 80's. The experiment was designed to be as simple as possible, in consideration for the age of the observers. We prepared 12 original speech segments, uttered by a skilled NHK announcer at a normal rate of 8.2 mora/s. These segments were categorized into four groups. Every segment in a group was con-

verted to the same speech rate in the linear mode, while the speech rates were mutually different among the groups. The rate was selected from four speech rates: 6, 7.2, 8.2, and 10 mora/s. Thus, each original speech segment was converted to just one speech rate.

Each of the 12 converted speech-segments was presented twice, following a start signal, through four loudspeakers in a large meeting room, where it had been confirmed in a preliminary test that the room itself caused no problem in intelligibility. The observers were instructed to choose one of four categories, consisting of 1 (unacceptable), 2 (acceptable), 3 (good), and 4 (very good). Valid answers were obtained from 279 observers.

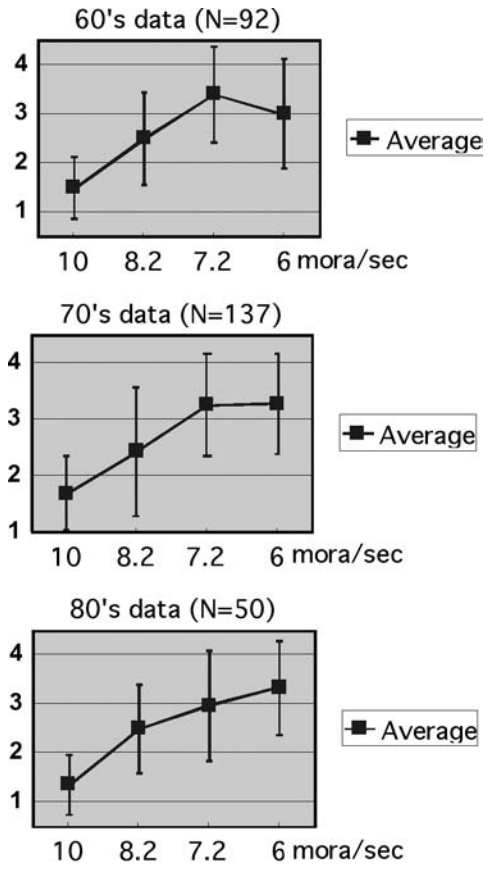

Figure 2. Evaluation scores as a function of the speech rates. The vertical lines indicate standard deviations of the distribution.

Results and discussion

The results are shown in Figure 2. Each point in the figure is the average score within the same speech rate group. The observers in their 60's judged the 7.2 mora/s rate as the best, while the subjects in their 80's preferred the 6.0 mora/s rate. The results show that most elderly observers preferred speech at a rate of no more than 7.2 mora/s, which is slower than normal.

REAL-TIME TRANSCRIPTION SYSTEM FOR NEWS PROGRAMS

The system

All TV programs are expected to be subtitled as soon as possible. Especially, subtitling for news programs is urgently required. The necessary target recognition rate for the news speech recognition system is no less than 95%, with a recognition delay of two seconds, for speech uttered by announcers in a news studio. The rate of 95% is significant, because the error rate of 5% is the upper limit of the ability to manually correct recognition errors in real-time.

The system consists of an acoustic analysis unit, a decoder, acoustic models, including

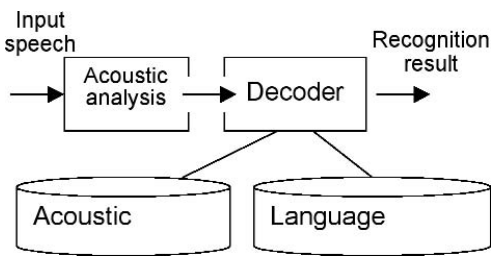

Figure 3. Block diagram of the continuous speech recognition system.

HMM (Hidden Markov Models), and language models, including word bigrams (connective probabilities between two adjacent words) and word trigrams (connective probabilities between three adjacent words). Figure 3 shows the block diagram of the system.

Acoustic Analysis

The procedure for acoustic analysis is as follows:

(a) Input speech is digitized at a sampling frequency of 16kHz and a quantization accuracy of 16 bits;
(b) The acoustic analysis unit executes short-time frequency analysis, using a Hamming Window with a 25ms analysis frame length and a 10ms frame period;
(c) The 12-dimensional MFCC (Mel Frequency Cepstrum Coefficient)[7] is calculated for each frame;
(d) The 12-dimensional delta-MFCC coefficient[8] is calculated by obtaining the slope of the least squares line against the train of coefficients of 5 frames, for each dimension of the MFCC.
(e) The 12-dimensional delta-delta-MFCC coefficient is calculated by performing the identical operation on the delta-MFCC coefficients.
(f) The logarithmic power of each frame, its delta and delta-delta are calculated

Finally, 39-dimensional acoustic parameters are obtained for each frame.

Acoustic models and language models

Gender-dependent, speaker-independent triphone-HMMs were used as the acoustic models[9]. The type of HMM was a continuously distributed HMM, using an 8-mixtured Gaussian distribution with diagonal covariance matrices. 42 Japanese phonemes were prepared for the system. The HMMs were learned using 100 phoneme-balanced news sentences, spoken by 45 NHK announcers, consisting of 24 male and 21 female announcers.

The language models were constructed using the following procedures:

(a) The news manuscripts were divided into morpheme units through automatic morpheme analysis, because Japanese sentences are not partitioned by word;
(b) The words defined by morpheme were selected in order of frequency of appearance in news manuscripts, and were registered in the word pronunciation dictionary;
(c) The bigrams and trigrams were calculated for the words registered in this dictionary;
(d) The Good-Turing method was used for Backoff smoothing[10].

Decoder

A two-pass decoder was adopted (see Figure 4) for selecting recognition candidates. In the

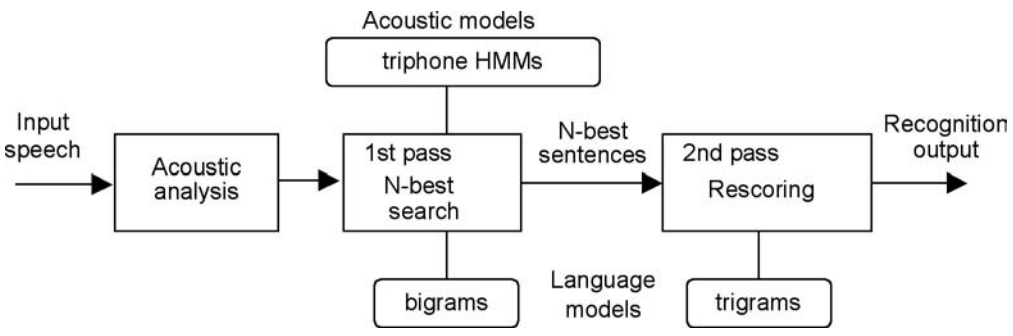

Figure 4. Detailed block diagram of the speech recognition system.

first pass, the word-dependent N-best search[11] is executed by a Viterbi beam search, using the word bigrams as the language model. The score of each candidate is calculated from the weighted acoustic (logarithmic probability) and language (logarithmic probability) scores. In the second pass, re-scoring is performed by re-calculation of each language score using word trigrams against the N-best sentences. The same weightings for the acoustic and language scores were used in the first and second pass.

NHK news speech database

To obtain excellent recognition performance in statistical speech recognition systems, it is indispensable to prepare a speech database with a large volume and high quality. We created a database of speech from actual news programs broadcast since 1996.

One of the weak points of a statistical language model is that appearance probabilities for recently appearing words are very low, despite their importance. Figure 5 indicates a new method for adapting the language models, using manuscripts prepared in advance that include almost all new words. Recent news manuscripts, duplicated 1000 times, are added to the news manuscripts accumulated over a number of years. As the result of this, even new words can be easily recognized by the system.

Evaluation of the system

Recognition tests were conducted to evaluate the performance of the broadcast news transcription system. The NHK news programs "News 7" and "Ohayo Nippon," broadcast from June 1 to June 7 in 1998, were used for the test set shown in Table 1.

For creating the acoustic models, the speech data selected from the database were broadcasts from one month in 1996, two months in 1997, two months in 1998 and 14 months from April 1999 to the end of May 2000. For learning the language models, manuscripts written from 1991 to the end of May 2000 were used.

We evaluated the results of recognition experiments by word recognition accuracy. The word recognition accuracy can be expressed as:where

$$A_w = (N_w - N_{subst} - N_{ins} - N_{del}) / N_w$$

A_w = word recognition accuracy
N_w = total number of words
N_{subst} = number of substituted words
N_{ins} = number of inserted words
N_{del} = number of deleted words

Results and discussion

The rightmost column of Table 1 shows the recognition results. Word recognition accura-

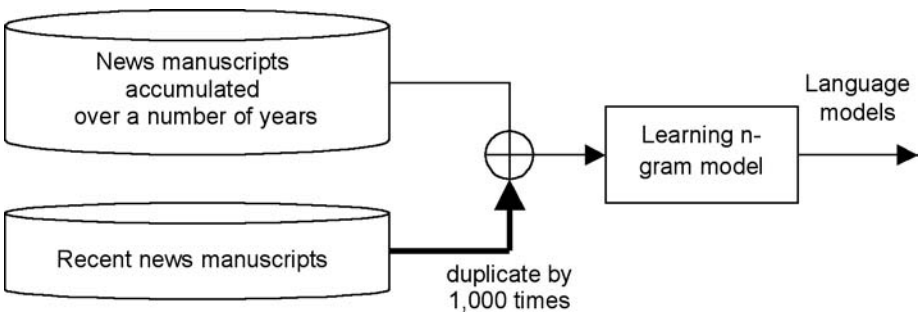

Figure 5. Method for adaptation of language models using recent news manuscripts.

Table 1: List of news programs for evaluation of the speech recognition system and its recognition results.

Condition	Number of sentences	Word accuracy %
Speech uttered by announcers in studio S/N = 55 dB	339	98.32
Spontaneous speech including commentary and conversational speech	160	78.05
Speech uttered by reporters in studio	30	87.86
Weather speech	340	77.21
Sports speech Clean S/N = 50 dB	139	88.09
Mixed with noise S/N=33 dB	254	72.15
Speech uttered by reporters outdoors	109	91.77
Speech uttered by announcers mixed with outdoor noise S/N= 30 dB	140	96.56

cy was 98.32% for the speech of studio announcers and 91.77% for the speech of reporters outside the studio. Of particular interest is the fact that a word recognition accuracy of 96.56% was obtained for the speech of studio announcers mixed with outdoor noise. A time of less than two seconds was the actual recognition time delay, defined as the time difference between the onset of the input speech and the start of the output of the recognition. It can be said that real-time recognition was realized. The necessary target recognition rate greater than 95%, with a recognition delay of 2 seconds or less, has been successfully reached for speech uttered by announcers in a studio for a news program.

This recognition system has already been installed in a live news studio since the end of March 2000, and the closed-captioning service on the NHK news programs "News7" and "News9" is now running well, although the captioning is restricted to speech uttered by announcers in the studio. The recognition rates for other speech in a news program, however, are less than 95%. This recognition rate may be improved by preparing various acoustic and language models, fitted to each type of speech categorized in Table 1.

CONCLUSION

The systems introduced in this paper are examples of applying speech processing to human-friendly interfaces. In order to evolve interfaces for elderly TV viewers, interfaces that are more intelligent should be developed. Recently, the therapeutic approach has become important in designing human-friendly interfaces. Various pet-robots, for example, have been developed to divert the elderly. The concept of agent TV has also been proposed. For such interfaces, the quality of human-friendship is expected to be studied.

References

1. Musiek E, Baran JA. Amplification and the Central Auditory Nervous System. In: Valente M, editor. Hearing Aids: Standards, Options, and Limitations. New York: Thieme Medical Publishers; 1996; pp 407-37.
2. Herold EW. Audio for the Elderly. J Audio Eng Soc 1988; 36:10.
3. Tosa T. Rapid speech in broadcasting. The NHK Monthly Report on Broadcast Research 1992:58-61.
4. Teranishi R. Critical Rate for Identification

and Information Capacity in Hearing System. J Acoust Soc Japan 1977; 33:136-143.
5. Nakamura A, Seiyama N, Ikezawa R, Takagi T, Miyasaka E. Real time voice speed converting system with small impairments. J Acoust Soc Japan 1994; 50:509-520.
6. Miyasaka E, Nakamura A, Seiyama N, Imai A, Takagi T. A New Approach to Compensate for Degeneration of Hearing Intelligibility in Elderly Listeners. The 100th Convention of the Audiological Engineering Society 1996:K-6.
7. Davis SB, Mermelstein P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Trans Acoust Speech Signal Proc 1980; 28:357-366.
8. Furui S. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. IEEE Trans Acoust Speech Signal Proc 1986; 34:52-9.
9. Schwartz R. Improved Hidden Markov Modeling Phonemes for Continuous Speech Recognition. Proceedings IEEE ICASSP-84. 1984;35:6.
10. Katz SM. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Trans Acoust Speech Signal Proc 1987; 35:400-1.
11. Schwartz R. A Comparison of Several Approximate Algorithms for Finding Multiple (N-best) Sentence Hypotheses. Proceedings IEEE ICASSP-91. 1991:701-4.