Feature

Speech Synthesis in Program Production and Research Trends

Speech synthesis is a very convenient means of conveying spoken information without the need for human labor. The most commonly used method of speech synthesis in program production compiles speech recorded during different tasks. Additionally, the text-to-speech (TTS) synthesis, which can synthesize speech from any text, can be used to deal with dialogue spoken with unemotional or unnatural voices for the sake of dramatic or comedic effect. This paper describes the above techniques and research in this field and presents application examples.

1. Introduction

Speech synthesis and speech recognition are essential technologies for creating user interfaces that make use of speech. Ideally, speech synthesis would automatically produce speech in a way that reflects the intent, emotions, and other characteristics of the user's speech. The convenience of conveying spoken information generated by machine without human intervention has led to speech synthesis being used throughout society. It is being used in a variety of circumstances to provide guidance to the public and to make announcements in train stations and on buses. Most systems synthesize speech from compilations of recorded sound; they can be thought of as a type of speech encoding system that compresses, stores, and plays back recorded speech. Despite its simplicity, synthesis from compilation is capable of producing high-quality speech approximating that of a natural, human voice, and as such, it has found use in a variety of fields including broadcasting. On the other hand, the text-to-speech (TTS) synthesis can generate speech from any text and hence is a general purpose method. Its range of application is expanding as a result of improving quality, thanks to technological advances.

This paper describes speech-synthesis techniques used in program production and research trends in this field. Section 2 describes task-limited, high-quality speech synthesis from compilations of recorded sound, section 3 describes TTS techniques with a focus on corpus-based speech synthesis[*1], and section 4 presents examples of using speech-synthesis techniques in program production.

2. Synthesis from Compilations of Recorded Sound[1)-3)]

Synthesis from compilations of recorded sound involves accessing stored recorded utterances (speech segments) in units of words, phrases, and even sentences, combining those segments according to the content to be synthesized, and inserting silent pauses of appropriate duration between the segments. Provided that the target tasks are limited in nature, this method can easily create high-quality synthesized speech on a level nearly the same as a human voice. It has consequently found widespread use in tasks requiring a limited vocabulary, such as making announcements in train stations, providing verbal guidance for operating equipment such as a telephone answering machine, and providing road and traffic information to drivers. One method of obtaining speech segments is to store speech waveforms directly by waveform encoding[*2]. Another method is to compress and store articulation[*3] synthesis filters expressing a different spectrum (frequency characteristics) for each phoneme[*4] and source waveforms to be input to those filters in parameter format. In the latter method, parameters can be manipulated to lengthen or shorten the duration of phonemes and to interpolate between changes in pitch or spectrum at concatenation points.

Figure 1 shows an example of converting the text of a weather report into speech. In this example, speech segments for "tokyoto" [Tokyo], "kita yori no kaze" [wind from the north], and "hare nochi kumori desho" [clear in the morning and cloudy later] are recorded beforehand as speech segments, and each segment is assigned a corresponding symbol such as 101, 202, or 303. Given an input symbol sequence (101, 202, 303 in this case), the speech-selection section reads out the segments corresponding to those symbols from the speech waveform database. The speech-concatenation section then connects the segments while inserting silent pauses of appropriate duration between them and outputs

[*1] A technique for synthesizing speech based on a database called a "corpus" consisting of speech data, texts of utterances, and power, pitch, and spectra of auditory information obtained by analyzing that data and text.

[*2] A method that directly encodes speech waveforms.

[*3] Functions of the vocal tract related to the shape and movement of the mouth when pronouncing certain sounds.

[*4] An abstract sound identified and recognized as having different pronunciations. In Japanese, they correspond to kana (the Japanese syllabary).

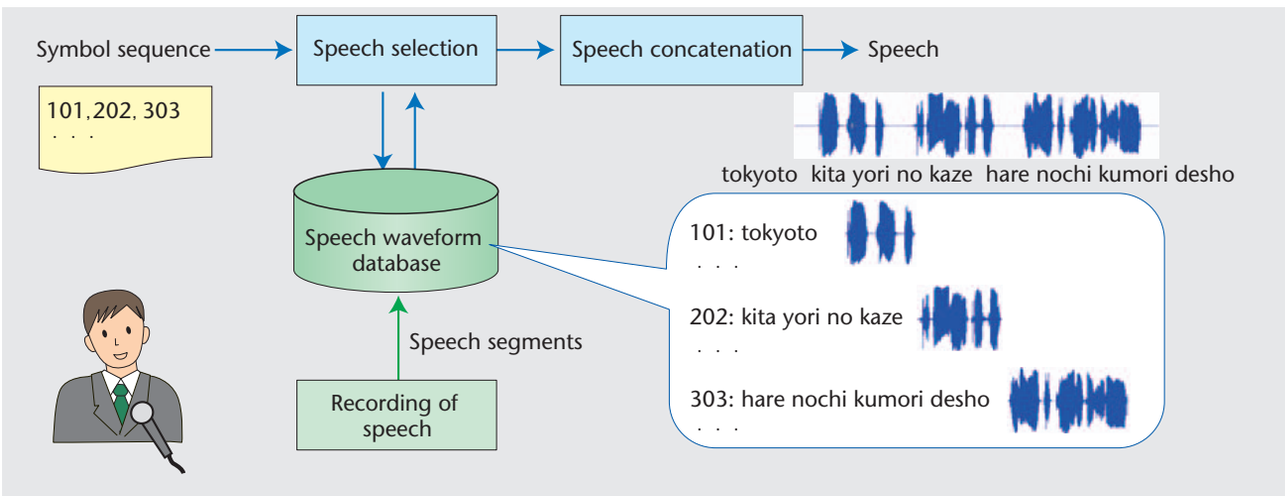

Figure 1: Synthesis from compilations of recorded sound

synthesized speech (here, the Japanese for "Tokyo, wind from the north, clear in the morning and cloudy later").

The more naturally that acoustic features (power, pitch, spectrum, and speech rate) flow in synthesized speech, the better the perceived quality of synthesized speech will be. For this reason, making the length of speech segments longer, as in the case of "hare nochi kumori desho" [clear in the morning and cloudy later], improves the quality of synthesized speech. At the same time, making segments longer reduces the number of sentences that can be synthesized using that segment, thereby making it necessary to record more speech segments overall. On the other hand, shortening segment units in the manner of "hare," "nochi," "kumori," and "desho" increases the number of sentences that can be synthesized by combining speech segments. This means there will be fewer segments that have to be recorded but synthesized speech will be of lower quality since the number of concatenation points increases. The usual approach to recording speech segments is to select units for which insertion of silent pauses of appropriate duration between the segments is enough to make the concatenation points sound natural.

Another way of using short units for speech segments is to modify acoustic features for the same word and prepare multiple versions of that speech segment as data.

In synthesis from compilations of recorded sound, new speech data has to be recorded when vocabulary is modified or added. To maintain the quality of synthesized speech in such a process, the same speaker who recorded the existing data should read for the recordings of the new data. That is, the acoustic features should be kept consistent in order to prevent unnatural connections.

3. TTS Synthesis[1)-3)]

The TTS synthesis automatically produces speech corresponding to the input text. This method uses speech segments in phoneme or syllable[*5] units smaller than words, which means that it can support a large vocabulary and synthesize any text, which is difficult for speech synthesis from compilations of recorded sound. On the other hand, TTS is inferior to synthesis from compilation

[*5] Unit of speech each consisting of a vowel or a vowel with surrounding consonants. In Japanese, they essentially correspond to kana (the Japanese syllabary).

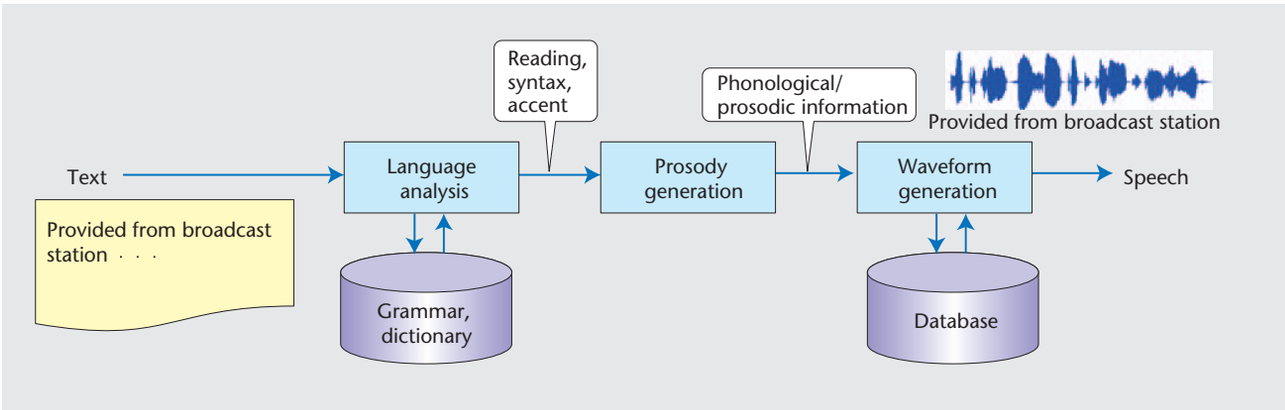

Figure 2: TTS synthesis

in terms of clear and natural synthesized speech, and it cannot easily make high-quality speech that can be used for broadcast. Despite this disadvantage, TTS is now equipped on personal computers (PCs) and on Web and e-mail applications to read out information for the benefit of visually impaired persons. The quality of TTS is improving as research progresses, and its applications are growing.

The TTS synthesis is outlined in Figure 2. Given the text, the language-analysis section begins the text-to-speech process by identifying words, providing their correct readings and pronunciations (phonological information), and supplying syntactical information such as dependency relationships, accent information, etc. Next, the prosody-generation section uses this information to generate information on natural intonation, the positions and lengths of pauses, etc., as it relates to the speaking of a sentence. Finally, the waveform-generation section uses phonological and prosodic information to generate speech waveforms and output synthesized speech. In applying TTS to actual broadcasts that require accurate information to be conveyed, the readings, syntactical dependencies, accents, etc., generated by the language-analysis section must be checked manually for errors before the speech is broadcast.

In the 1990s, rule-based speech synthesis became practical for a variety of situations. This method generates speech waveforms by controlling previously extracted and stored articulation synthesis filters and source-waveform parameters on the basis of acoustic and linguistic knowledge. However, the approach taken back then was based on empirical knowledge or craftsmanship-like know-how, which made it difficult to automatically construct a system. In addition, the speech so synthesized was not sufficiently natural. An alternative to rule-based speech synthesis is to accumulate a large amount of speech data together with its spectral and prosodic information as a speech corpus and to generate optimal speech based on appropriate evaluation measures. Research on corpus-based speech synthesis has been progressing. Compared with rule-based synthesis, corpus-based synthesis can synthesize more natural-sounding speech and can be used to automatically construct a speech-synthesis system. However, a considerable amount of speech data is needed to synthesize high-quality speech.

Waveform generation in corpus-based speech synthesis can be broadly divided into two methods: the unit-selection based synthesis that connects speech segments in appropriate synthesis units selected from the speech corpus and the statistical-model based synthesis that generates speech waveforms using parameters generated from a statistical model trained beforehand in terms of acoustic features. These two waveform-generation methods are described below.

3.1 Unit-selection Based Synthesis[4),5)]

In this method, there are no restrictions on the units of speech synthesis. Appropriate units are automatically selected from a speech corpus by using evaluation measures that provide criteria for selecting speech segments. Speech waveforms are then concatenated into synthesized speech. The process of synthesizing speech through waveform processing is called synthesis by concatenation and signal-processing modification, whereas synthesizing speech without any waveform processing is called concatenative synthesis. Singing-synthesis software has recently become popular, and some of this software uses synthesis by concatenation and signal-processing modification[6)]. This software inputs lyrics together with musical notes, thereby making it unnecessary to automatically generate prosodic information such as accents and intonation. The somewhat unemotional, unnatural aspect of the speech synthesized with this method is considered by users to have a personality of its own, and the singing it generates can be purely enjoyed as music produced from synthesized speech.

Figure 3 outlines the unit-selection based synthesis. Once phonological and prosodic information is input to the waveform-generation section, the unit-selection

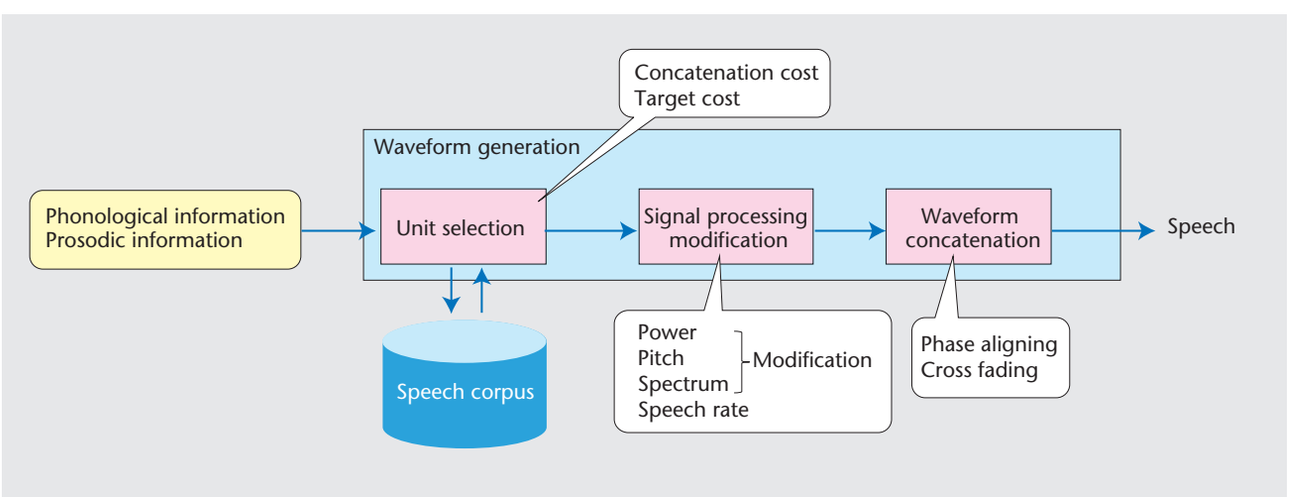

Figure 3: Unit-selection based synthesis

section makes up a list of synthesis-unit candidates (phoneme sequences). It then searches inside the speech corpus for the most optimal combination of synthesis-unit candidates corresponding to this list. This search is performed through dynamic programming[*6] using evaluation measures such as the fit between the extraction environment and usage environment, the degree of continuity (concatenation cost) when concatenating synthesis-unit candidates, and matching of generated prosody (target cost). Next, taking as input the synthesis units selected in the above way, the Signal-processing-modification section makes changes to acoustic features such as power, pitch, spectrum, and speech rate to reduce discontinuities when concatenating synthesis units and minimize mismatches in the generated prosody. Here, it is better not to change the original, natural waveform if at all possible in order to synthesize high-quality speech. Finally, the waveform-concatenation section aligns the waveform phases of the synthesis units modified by the Signal-processing-modification section, applies cross fading[*7] when concatenating the waveforms to prevent discontinuities, and outputs synthesized speech.

Compared with rule-based synthesis, the unit-selection based synthesis must store a large amount of waveform data but it can synthesize clear and natural speech by making only slight changes to the original speech in the corpus. This is why this method is currently being used in commercial speech-synthesis software. Significant variation in the quality of synthesized speech can occur depending on the input text, and though high quality can be achieved if speech units can be skillfully concatenated, discontinuities will occur and quality will drop if there are no speech units in the corpus that can be neatly concatenated. Consequently, whether high-quality speech can be synthesized depends on whether the speech corpus can cover any text.

3.2 Statistical-model Based Synthesis[7)-9)]

The statistical-model based synthesis first models the acoustic features of synthesis units using a statistical model such as the hidden Markov model (HMM)[*8] and generates parameters from this model. Then, based on these parameters, the process controls the articulation synthesis filter and source waveforms and synthesizes speech.

Figure 4 outlines the statistical-model based synthesis. The parameter-extraction section in the pre-training section extracts static and dynamic parameters beforehand for the spectral and excitation parameters for each context-dependent synthesis unit from the speech corpus. Here, synthesis units are determined by considering a variety of linguistic features such as the preceding and subsequent phonological environments, accents, and parts of speech. Next, the statistical-model-training section performs machine training, treating the static and dynamic parameters for spectral and excitation parameters extracted in the previous step as one item of data. It then models the synthesis units and creates a statistical-model database. When phonological information and prosodic information is given to the waveform-generation section, the statistical-model-

[*6] An efficient technique for solving an optimization problem by dividing it into several parts, solving each part in order, and eliminating those parts for which an optimal solution better than those so far determined cannot be obtained.

[*7] A technique for overlapping and concatenating sounds by gradually diminishing the end of the previous sound and gradually boosting the start of the following sound so as to make a smooth connection between them.

[*8] A technique that estimates unknown, hidden parameters from observable information by assuming a Markov process in which the future state of a system is dependent only on its present state, i.e., independent of past states.

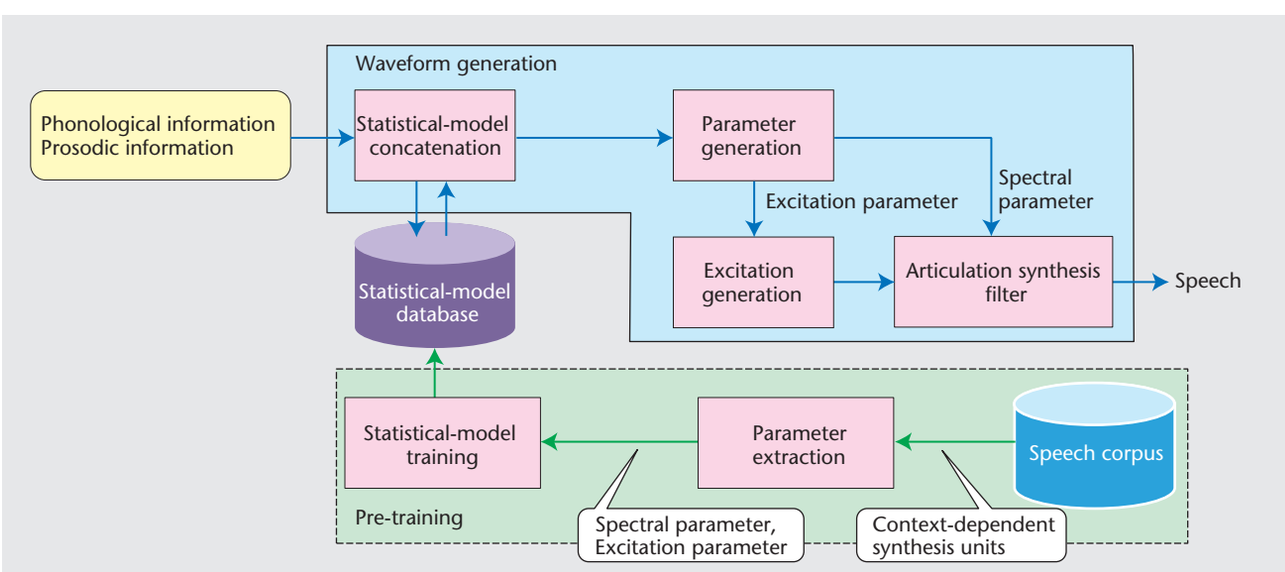

Figure 4: Statistical-model based synthesis

concatenation section creates a sequence of context-dependent synthesis units, selects the statistical model corresponding to each synthesis unit, and concatenates the models. After this, the parameter-generation section generates spectral and excitation parameters needed for speech synthesis from the statistical models concatenated in units of sentences. The articulation-synthesis-filter section then uses these spectral parameters to configure filters for expressing spectrums that differ for each phoneme. The excitation-generation section, in turn, uses the excitation parameters to generate excitation waveforms and input them to the articulation-synthesis-filter section in periodic units for expressing prosody, that is, changes in pitch. The output from the articulation-synthesis-filter section is synthesized speech.

Compared with the unit-selection based method, the quality of speech synthesized by the statistical-model based method is often inferior. Despite this, however, the method can synthesize smooth speech with highly natural prosody from a smaller amount of data. The statistical-model based method is therefore considered to be effective for synthesizing speech with diverse quality and speaking patterns. It is becoming a dominant subject in speech-synthesis research.

4. Use in Program Production[3)]

Speech synthesis has been used in broadcasting since the 1990s. Moreover, progress in computers and communications and advances in speech-synthesis technology have gone hand in hand during this time. Broadcast stations consider speech synthesis to be a highly useful technology since it enables information to be conveyed by synthesized speech early in the morning or late at night when anchorpersons may not be present at the studio.

Synthesized speech is most commonly used at broadcast stations for automatically broadcasting weather reports[10)-13)]. It is also used for broadcasting traffic information[14)] and earthquake and tsunami bulletins[15),16)]. These systems are mostly compilations of recorded sound, in light of their need for high-quality speech.

The TTS method has also been used in broadcasting for narrating variety programs, and its somewhat unemotional, unnatural synthesized speech has come to thought of as having a personality in its own right. This method is being used by regional commercial broadcasters to deliver local weather reports and for reading out stock prices on shortwave radio broadcasts.

NHK uses a TTS technique based on the unit-selection based method. For example, TTS has been used for "News at Anytime"in which the news section of the NHK home page are readout in Japanese and for"Multilingual Weather Report" in which weather reports are read out in Japanese, English, Chinese, and Korean as part of the digital radio trials. NHK has also developed a speech-synthesis system that combines a waveform-concatenation type of numerical speech synthesis with synthesis from compilations[17)], and it has used this system since fiscal year 2010 to read out stock prices for "Stock Market Report" on NHK Radio 2.

5. Conclusion

This paper described speech-synthesis techniques used in the production of television programs and discussed research trends and examples of those techniques in actual program production. The techniques discussed here can be broadly divided into synthesis from compilations of recorded sound and text-to-speech synthesis, and each method has its advantages and disadvantages. Since the quality demanded of synthesized speech depends on the application, the individual characteristics of each technique should be put to good use for different needs. For programs that aim to convey fixed-form information as in weather reports, importance should be placed on correct read out of information and on voice quality, clarity, and naturalness comparable to that of an anchorperson; synthesis from compilations of recorded sound is generally considered to be the optimal technique at present for such programs. The quality of text-to-speech synthesis has been improving, thanks to progress on corpus-based speech synthesis (unit-selection and statistical-model methods), and if this approach can reach a level at which it can express a variety of speaking styles with various levels of voice quality and emotion, its use should expand beyond that of reading out of information to other uses such as generating voices for animated characters. Additionally, the ongoing convergence of broadcasting and communications will be an engine for the creation of new and diverse services. That means we can expect corpus-based speech synthesis to find its way into receivers, personal computers, mobile phones, and smartphones in the near future.

(Nobumasa Seiyama)

References

1) S. Furui: Speech Information Processing, Morikita Publishing, pp. 67-78 (1998) (in Japanese)
2) S. Itahashi: Speech Engineering, Morikita Publishing, pp. 150-176 (2005) (in Japanese)
3) N. Yagi: Image Media Technology, Ohmsha, pp. 201-211 (2008) (in Japanese)
4) Y. Sagisaka: "Technology Trends in Corpus-based Speech Synthesis [I]: The Past, Present, and Future of Corpus-based Speech Synthesis," Journal of the IEICE, Vol. 87, No. 1, pp. 64-69 (2004) (in Japanese)
5) M. Abe: "Technology Trends in Corpus-based Speech Synthesis [II]: Explanations Using Speech Synthesis Units," Journal of the IEICE, Vol. 87, No. 2, pp. 129-134 (2004) (in Japanese)
6) H. Kenmochi: "Recent trend of singing synthesis: Technology that supports 'Hatsune Miku'," Journal of the Acoustical Society of Japan, Vol. 67, No. 1, pp. 46-50 (2011) (in Japanese)
7) T. Kobayashi and K. Tokuda: "Technology Trends in Corpus-based Speech Synthesis [IV]: HMM-Based

Speech Synthesis," Journal of the IEICE, Vol. 87, No. 4, pp. 322-327 (2004) (in Japanese)
8) T. Kobayashi: "Foreword to the special issue on recent progress of speech synthesis research," Journal of the Acoustical Society of Japan, Vol. 67, No. 1, pp. 15-16 (2011) (in Japanese)
9) K. Tokuda: "Recent advances in statistical parametric speech synthesis," Journal of the Acoustical Society of Japan, Vol. 67, No. 1, pp. 17-22 (2011) (in Japanese)
10) Y. Hiraoka and H. Uchiyama: "The Program Production Technique on 'Early Morning and Night Weather Forecast'," Journal of the Institute of Television Engineers of Japan, Vol. 41, No. 8, pp. 742-743 (1987) (in Japanese)
11) H. Kitahama, J. Imakawa and H. Miura: "Weather Reporting System with Automatic Speech Synthesis, Proceedings of the 31st Technical Report Conference of Commercial Broadcasters, 11 (1994) (in Japanese)
12) F. Sawaguchi: "A Speech Synthesizer for Weather Information System," ITE Technical Report, 21, 53, pp. 25-30 (1997) (in Japanese)
13) H. Adachi: "Automatic Weather Announcements System," Proceedings of the 57th NHK Forum on Technology of Broadcasting, 14 (2004) (in Japanese)
14) T. Kawakami and S. Shimono: "Fully Automatic Road and Traffic Report System using CG and Speech Synthesis," Proceedings of the 39th Technical Report Conference of Commercial Broadcasters, 5 (2002) (in Japanese)
15) Y. Miyasaka: "Development of automatic sending system of earthquake and tidal wave radio quick report," Proceedings of the 50th NHK Forum on Technology of Broadcasting, 8 (1997) (in Japanese)
16) S. Miura and T. Ito: "PC-based Earthquake Bulletin System with Speech," Proceedings of the 38th Technical Report Conference of Commercial Broadcasters, 6 (2001) (in Japanese)
17) H. Segi, N. Seiyama, R. Tako, T. Takagi, S. Oode, A. Imai, M. Nishiwaki and R. Koyama: "Developing Speech Synthesis System for Stock-price Bulletins and Trial Use in Digital Terrestrial Radio," Journal of the Institute of Image Information and Television Engineers, Vol. 62, No. 1, pp. 69-76 (2008) (in Japanese)