A review of paralinguistic information processing for natural speech communication

Yoichi Yamashita*

College of Information Science and Engineering, Ritsumeikan University, Noji-higashi 1–1–1, Kusatsu, 525–8577 Japan

Abstract: Speech conveys not only linguistic information but also supplemental information that is not inferable from written language, such as attitude, speaking style, intention, emotion, mental state, and so on, and is called para- or non-linguistic information. This type of information plays important roles for smooth and natural communication through spoken language. This paper reviews recognition and synthesis techniques for speech communication focusing on emotion and emphasis as well as corpora that are dispensable to development of current speech technologies.

Keywords: Man-machine interaction, Emotion, Emphasis, Expressive speech, Corpus

PACS number: 43.72.Ja, 43.72.Kb, 43.72.Ne, 43.70.Fq [doi:10.1250/ast.34.73]

1. INTRODUCTION

Speech is a useful and efficient communication media in a daily life between human beings and it is expected to be utilized in man-machine interaction. Speech conveys several types of information and Fujisaki classified information expressed by speech into three categories; linguistic, paralinguistic, and nonlinguistic information [1]. Linguistic information is defined as “the symbolic information that is represented by a set of discrete symbols” and it can be transcribed in written language. Paralinguistic information is “added by the speaker to modify and supplement the linguistic information.” It can be consciously controlled by the speaker and includes intentions, attitudes, emphasis, speaking styles, and so on. Nonlinguistic information, which is sometimes referred as the extra-linguistic information [2], concerns “such factors as the age, gender, idiosyncracy, and physical and emotional states of the speaker, and so on, and can not generally be controlled by the speaker.” However, “a speaker can control the way of speaking intentionally to convey an emotion” even if he is not a professional actor. In this paper, the term of paralinguistic is used as an extended concept that is concerned with a part of nonlinguistic of the above definition and includes emotional states of a speaker.

One of benefits of communication through spoken language is that parts of information that we want to convey are easily and naturally encoded in speech with paralinguistic information, whereas we write an e-mail carefully to avoid miscommunication because written language does not convey paralinguistic information. Many studies of paralinguistic information processing have been conducted to realize natural man-machine interaction through spoken language paying attention to recognition and synthesis of paralinguistic information [3–7]. After “ISCA ITRW on speech and emotion” was held in 2000 [8], contest-based special sessions of paralinguistic information processing sharing the same data was organized as “INTERSPEECH 2009 Emotion Challenge” and “INTERSPEECH 2010 Paralinguistic Challenge” in 2009 and 2010 [9,10], a special issue of “Sensing Emotion and Affect” was described on the journal of Speech Communication [11], and other activities are found at [12].

This paper reviews recognition and synthesis techniques for speech communication focusing on emotion and emphasis as well as corpora that are dispensable to development of current speech technologies.

2. CORPUS

Most recent studies of spoken language processing are based on statistical methods that require a large size of data sufficient to train mathematical models. A great advantage of statistical methods over rule-based methods is reproducibility that enables anybody to get the same results if he or she uses the same data and modeling, while it is difficult for researchers to share all heuristic rules in the rule-based methods. On the other hand, a disadvantage is that it is not easy to build a corpus by collecting and annotating large amount of data. Corpora are indispensable not only to training and evaluation of statistical modeling but also to
investigation of acoustic and linguistic characteristics of speech, relationship between acoustic and linguistic information, and so on. How to build a corpus is a crucial issue in current speech research based on statistical methods. Many speech corpora have been developed for research related to the paralinguistic information [13–17]. Many existing speech corpora are listed in the literatures [6,14,17].

2.1. Naturalness

One of considerable points of developing speech corpora for paralinguistic research is naturalness of paralinguistic information of speech. To record speech with rich paralinguistic information, such as emotional speech, we may ask speakers to utter speech with required emotions using scripts. Trained or professional speakers can act and utter emotional speech intentionally and such acted speech is easy to control an emotion conveyed in the speech. However, it is uttered in an artificial situation and is characterized by different properties from a genuine natural emotion.

On the other hand, conversations in a daily life and dialogs in a designed task are recored to collect natural emotional speech [13,17]. In the recording natural speech, it is difficult to control emotions in speech and natural speech sometimes lacks phonetic and prosodic valances in a corpus. Nevertheless, there is a trend of development of speech corpora containing natural emotional speech because the realistic applications of recognition or synthesis of paralinguistic information should be developed based on natural speech.

2.2. Annotation

There are two types of schemes for annotating speech data with paralinguistic information; categorical and dimensional. In the categorical annotation, for example, a set of discrete labels, such as anger, sadness, happiness, fear, disgust, and so on, is defined for emotion annotation. An appropriate label is assigned to utterances or segments of speech, and speech data is sometimes annotated with scaling as well as the choice of a label. In the dimensional annotation, a set of fundamental dimensions that spans a space of paralinguistic information. Each utterance or segment is assigned with scales of the dimensions, and is expressed as a point in the paralinguistic space. Although most studies of recognition and synthesis of paralinguistic information use the labels of the categorical annotation, the dimensional annotation is suitable to describe paralinguistic information in natural speech because the paralinguistic state intrinsically has the property of graduation and uncertainty. Two major dimensions for the dimensional annotation of emotions are activation and evaluation [18]. The activation is a dimension for describing active and passive and the evaluation is for positive or negative. The former is sometimes called as arousal and the latter as valence or pleasantness. Another third dimension that is called strength or power sometimes describes attention or rejection [19,20].

Mori et al. developed a Japanese corpus for studying paralinguistic information, named the Utsunomiya University Spoken Dialog Database for Paralinguistic Information Studies (UU database) [17,21]. The UU database contains utterances in dialogs of “4-fame cartoon sorting task.” Each utterance is annotated on a 7-point scale for 6 dimensions, which include 4 extensive dimensions as well as the 2 major dimensions, by evaluating the perceived emotional states of the speakers. This corpus has a remarkable feature of public availability and it can be accessed through NII (National Institute of Informatics) [22].

Paralinguistic labels are obvious for acted speech that speakers are asked to utter with requirement. On the other hand, annotating speech data with paralinguistic information for natural speech is a very time-consuming task. It is expected to develop automatic annotation techniques that yield reliable and consistent labels of paralinguistic information [23] as well as prosodic information [24,25].

3. RECOGNITION OF PARALINGUISTIC INFORMATION

Recognition of paralinguistic information in man-machine interaction can realize an insightful interface system by understanding the mental state of human users as well as lexical contents of the utterance. However, some reasons make it difficult to recognize paralinguistic information in speech, such emotion, emphasis, attitude, and so on. First, the way of expressing paralinguistic information is much dependent on the speaker and there are many prosodic and lexical variations. Second, the mental state of speakers is expressed not only by speech but also by other modalities, such as facial expressions and motion. Third, there may be more than one paralinguistic information in an utterance and the boundaries between emotions are ambiguous.

3.1. Emotion Recognition

The first problem of emotion recognition is categorization of emotions. Psychological studies have described the organization of emotion types and basic emotions [26]. Most studies that are concerned with emotion recognition use some primary emotions. Cowie et al. list anger, disgust, fear, joy, sadness, and surprise as the “archetypal” emotions that are the most obvious and distinct emotions [18]. Furthermore, it is suggested that simple categorical classification is not sufficient for emotion recognition and strength of the emotion should be estimated to leverage
paralinguistic information in speech communication [27]. The “Affect Sub-Challenge” in “INTERSPEECH 2010 Paralinguistic Challenge” dealt with prediction of “Level of Interest” of speakers [10], and it is one of hot topics of paralinguistic recognition [28]. Nose et al. proposed a method of estimating the degree of several emotions based on an emotional speech synthesis technique [29]. Each category of the emotion is often represented by stochastic modeling techniques like HMM and the most probable category is selected. On the other hand, a few studies recognize the emotion on the basis of some dimensional properties like activation and evaluation using multipletage classification [30].

The second problem is what emotional speech is used. In recording acted emotional speech, speakers are asked to utter the same sentence with required different emotions. In general, the emotion is much dependent on lexical contents of the utterance, but there are no relationship between an emotion and lexical contents in acted speech. Speakers may exaggerate utterance emotionally to make emotions more distinctive. An emotion in genuine natural speech is less distinctive and is sometimes mixed with other ones. Although acted speech is easy to control and is useful to investigate characteristics of emotions, it is not appropriate to develop realistic applications of speech communication.

The third problem is how results of emotion recognition are used. There are many possible applications for speech communication, but many studies of emotion recognition are aiming to develop a fundamental technique and are not oriented to any specific application. One of applications of emotion recognition is evaluation of user utterances in speech communication with a robot. Fujie et al. describe a robot that recognizes a user’s attitude as positive or negative [31]. Another unique application-oriented study is the anger detection for utterances in dialog for voice portal systems or contact center calls [32–34]. This technique is useful to measure the quality of dialogs and to detect potential problems caused by an unsatisfactory interaction.

3.2. Emphasis Detection

The speaker’s attention varies according to the context in a dialog even if lexical content of an utterance is the same. Words or phrases that a speaker pays attention have acoustical prominence or emphasis. Terken defines prominence as “the property by which linguistic units are perceived as standing out from their environment” [35] although there may be several viewpoints in defining the term of prominence. In spoken dialog systems, emphasis detection is an important technique because emphatic linguistic units can be key information for identifying attention or intention of the user.

Several studies concerned with emphasis detection have been conducted [36–39] and there is a rough agreement that duration, fundamental frequency ($F_0$), and intensity are related to emphasis in speech. Some studies proposed the usage of lexical or syntactic information for emphasis detection [37,38]. Most studies evaluate proposed methods using isolated read utterances. It is necessary to extend research targets to dialog spontaneous speech to incorporate emphasis detection into speech communication systems. In [40], emphasis detection is utilized for generating a feedback in tutoring communication with a robot.

4. SYNTHESIS OF EXPRESSIVE SPEECH

While intensive researches have improved intelligibility and naturalness of synthetic speech, variability is another important topic in speech synthesis to generate speech with various personality and paralinguistic information. Many studies have been conducted for synthesizing expressive speech which conveys rich paralinguistic information such as emotion, speaking style, and contrastive emphasis. This section reviews studies of expressive speech synthesis focusing on emotional speech synthesis (ESS) and generation of emphasis.

4.1. Synthesis of Emotional Speech

4.1.1. Rule-based approach

Emotions are expressed in speech by acoustic correlates of prosodic parameters, $F_0$, duration, and energy, and voice quality. Many literatures have describes acoustical characteristics of emotions [4,19,41]. Murray et al. describe about “anger,” “happiness,” “sadness,” “fear,” “disgust” as the primary emotions, and other several emotions as the secondary emotions [19]. Common understandings are found in studies of acoustics characteristics of emotions, for example, “anger” and “happiness” increase $F_0$ and “sadness” lengthens the duration.

In early studies of ESS, prosodic parameters are modified by heuristic rules that experts manually describe based on analysis results [42–46]. While the rule-based approach does not require large-scale speech corpus, the performance of ESS is dependent on a skill of the expert and it is difficult to improve proposed methods by other researchers.

4.1.2. Corpus-based approach

If the corpus of emotional speech data is available, we can develop ESS by using models which are trained from the corpus. There is a trend of shifting from the rule-based approach into the corpus-based approach in ESS as well as usual TTS for reading with neutral emotion. The corpus-based approach is divided into two categories, unit selection (US) and statistical parametric speech synthesis (SPSS).
The US-based synthesis system is implemented on the basis of concatenation of waveform inventories that are stored in a corpus. The US-based ESS requires a large scale corpus of emotional speech, and emotional property of synthetic speech is colored by acoustic correlates represented by inventories in a corpus [2,47–51]. The quality of ESS is dependent on the corpus and how to select inventories from the corpus. Iida et al. describe an ESS system for three emotions, “anger,” “joy,” and “sadness,” by constructing different corpora for each emotion [48]. This approach is simple but it needs large costs for constructing corpora. Pitrelli et al. proposed another approach of the US-based ESS by building a corpus that mixed several emotions and by introducing the emotion as a feature for inventory selection [49]. Moriyama et al. proposed an idea for representing relationship among $F_0$, energy, and duration using PCA on a subspace in ESS for Japanese words [50,51].

The SPSS systems which use Hidden Markov Model (HMM) as a statistical modeling technique of time sequences of parameters are called HMM-based speech synthesis [52,53]. Many techniques for HMM-based speech synthesis have been intensively proposed to realize flexible and high-quality speech synthesis, including ESS [54,55]. One of problems in ESS is that only small size of emotional speech corpus is often available or it not easy to collect large amount of emotional speech. An advantage of HMM-based speech synthesis is easy adaptation using small amount of data. HMM works as a generator of parameter sequences in HMM-based speech synthesis for spectrum, which are often represented with MFCC, and $F_0$. Outputs generated from HMMs can be modified by changing model parameters of the HMMs according to adaptation data. Yamagishi et al. describes MLLR adaptation from reading style to emotional speech for HMM-based speech synthesis [55].

4.1.3. Description of emotion

A natural emotion is not a binary feature of speech and its degree changes continuously. In addition, speech does not convey only one emotion but sometimes expresses multiple emotions and speaking styles simultaneously. To synthesis natural emotional speech, the input description into ESS should be a flexible representation of emotions that can describe the degree of emotions to be expressed [20]. Tachibana et al. proposed a method of mixing two emotions and expressing an intermediate emotion using parameter interpolation for HMM-based speech synthesis [56]. Nose et al. successfully introduced the style vector that describes the degree of emotions and speaking styles into a multiple-regression model of HMM-based speech synthesis [57], and also proposed a technique for speaker adaptation for this method [58].

For a corpus used for model training of SPSS, there can be two types of schemes for emotion annotation, intended emotion and perceived emotion. Speakers utters some sentences with required emotions for collecting emotional speech used for studies of ESS. The intended emotion is a emotion that a speaker is asked to express. On the other hand, the perceived emotion is given by subjective evaluation in listening test. Some studies describe that the perceived emotion improves the performance of ESS [59,60].

4.1.4. Evaluation

In most studies of ESS, quality of emotional expressivity is evaluated sentence by sentences by a forced choice test, that asks listening subjects to select one category of the emotion. In general, emotional speech output is not provided a user without context of preceding utterances in applications like reading books and a spoken dialog system. The forced choice test is not sufficient to evaluation when ESS is used in applications. In addition, there is an issue of selecting sentence contents that are used for listening evaluation. Emotionally neutral sentences are often used for evaluation. However, emotion is highly correlated with lexical contents as well as acoustic characteristics. It is necessary to evaluate emotional speech considering lexical contents, especially naturalness of emotional speech with context. Furthermore, Murray suggested that “emotionally undetermined” sentences are used for evaluation [44]. For example, “no one has telephoned me today” is emotionally neutral, but can satisfactorily take different emotions.

It is said that anger is similar to happiness for acoustic characteristics, such as high $F_0$ and large power. It is not necessary that these emotions are distinguished by only acoustic characteristics in speech for applications of speech communication because they often appears with different lexical expression and discourse context. Emotional speech should be evaluated in terms of naturalness in context as well as expressivity of isolated sentences.

4.2. Generation of Emphasis

For speech synthesis of reading English sentences, it is important to assign prominence, such as the location of a pitch accent on word or syllable. Assignment of standard pitch accents is not sufficient to synthesis natural speech with context, such discourse-level information [61]. Prediction of emphatic accent, which is called as ‘emphasis’ in this paper, improves naturalness of synthetic speech conveying speaker’s intention or mental state. Although Japanese speech synthesis considers accent types instead of pitch accents, it is also important to generate appropriate emphasis on synthetic speech with context [62]. For other language, relationship between prosodic information and emphasis was investigated [63].

The US-based speech synthesis expects that a speech corpus of inventories stores all of acoustic varieties that
may be used. How to record speech with emphasis is a key problem of emphasis generation as well as a mechanism of selecting inventories that introduces emphasis-related features [49,64].

In HMM-based speech synthesis, a corpus containing sentences with emphasis is used to train HMM models, and emphasis-related features are added to the feature set that is used for context clustering to merge similar models [23,65]. Maeno et al. proposed a method of automatic annotation of emphatic accent phrases [23].

### 4.3. Toward Application Development

Another problem of speech synthesis systems that express emotion or emphasis is that how to use the systems. For example, an application of reading books has to extract emotion to be expressed or emphatic words from written text. In dialog systems, a speech interface should reason an expressivity appropriate to a current user to communicate a spoken message efficiently. One solution is that a dialog control module predicts emotions or emphatic words based on dialog context [66,67].

In general, the US-based speech synthesis generates better voice quality but requires more speech data over the SPSS-based speech synthesis. On the other hand, the SPSS-based speech synthesis has advantages of flexibility and small foot prints. The best technique of speech synthesis is much dependent on applications and speech data available to develop systems. If ESS is required for various speakers’ voices, SPSS may provides a better solution than US. If the personality of a speech interface can be fixed, US can provide high quality of emotional and emphatic synthetic speech by collecting various speech data of a target speaker.

## 5. CONCLUSIONS

Several services of the speech interface on mobile phones have been put into real-world use [68,69]. They successfully provide users with an easy input method into mobile small devices using speech recognition. Development of new technologies are necessary to make the current speech interface more sophisticated or to realize more natural speech communication for other tasks. One direction of new technologies is recognition and generation of paralinguistic information. It can enable to understand user’s mind that is not explicitly contained in linguistic expression and to generate expressive spoken messages that are transfered without confusion. Appearance of comfortable and enjoyable speech communication systems based on recognition and synthesis of paralinguistic information is expected in the near future. We express paralinguistic information in speech using different ways of unconscious manipulation of acoustic correlates. Adaptation or customization to an individual user is also one of important issues of paralinguistic information processing.

## REFERENCES

[1] H. Fujiyki, “Information, prosody, and modeling —with emphasis on tonal features of speech—,” Proc. Speech Prosody 2004, Nara, pp. 1–10 (2004).
[2] N. Campbell, “Developments in corpus-based speech synthesis: Approaching natural conversational speech,” IEICE Trans. Inf. Syst., E88-D, 376–383 (2005).
[3] M. Schröder, “Emotional speech synthesis: A review,” Proc. EUROspeech 2001, Aalborg, pp. 561–564 (2001).
[4] D. Erickson, “Expressive speech: Production, perception and application to speech synthesis,” Acoust. Sci. & Tech., 26, 317–325 (2005).
[5] N. Campbell, W. Mamza, H. Höge, J. Tao and G. Bailly, “Special section on expressive speech synthesis,” IEEE Trans. Audio Speech Lang. Process., 14, 1097–1098 (2006).
[6] M. E. Ayadi, M. S. Kamel and F. Karay, “Survey on speech emotion recognition: Features, classification schemes, and database,” Pattern Recognit., 44, 572–587 (2011).
[7] B. Schuller, A. Batliner, S. Steidl and D. Seppi, “Recognizing realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” Speech Commun., 53, 1062–1087 (2011).
[8] ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, http://www.isca-speech.org/archive_open/speech_emotion/
[9] B. Schuller, S. Steidl and A. Batliner, “The INTERSPEECH 2009 Emotion Challenge,” Proc. INTERSPEECH 2009, Brighton, pp. 312–315 (2009). http://emotion-research.net/sigs/speech-sig/emotion-challenge
[10] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller and S. Narayanan, “The INTERSPEECH 2010 Paralinguistic Challenge,” Proc. INTERSPEECH 2010, Chiba, pp. 2758–2761 (2010). http://emotion-research.net/sigs/speech-sig/paralinguistic-challenge
[11] B. Schuller, A. Batliner and S. Steidl, “Introduction to the special issue on sensing emotion and affect—Facing realism in speech processing,” Speech Commun., 53, 1059–1061 (2011).
[12] http://emotion-research.net/sigs/speech-sig
[13] N. Campbell, “Speech & expression; The value of a longitudinal corpus,” Proc. Language Resource and Evaluation Conf., Lisbon, pp. 183–186 (2004).
[14] E. Douglas-Cowie, N. Campbell, R. Cowie and P. Roach, “Emotional speech: Towards a new generation of databases,” Speech Commun., 40, 33–60 (2003).
[15] F. Burkhardt, A. Paeschke, M. Rolles, W. Sendlmair and B. Weiss, “A database of German emotional speech,” Proc. INTERSPEECH 2005, Lisbon, pp. 1517–1520 (2005).
[16] M. Avanzi, A. C. Simon, J.-P. Goldman and A. Auchlin, “C-PROM: An annotated corpus for French prominence study,” Proc. Speech Prosody 2010, Chicago, 4 pp. (2010).
[17] H. Mori, T. Satake, M. Nakamura and H. Kasuya, “Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics,” Speech Commun., 53, 36–50 (2011).
[18] R. Cowie, E. D. Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor, “Emotion recognition in human-computer interaction,” IEEE Signal Process. Mag., 18, 32–80 (2001).
[19] I. R. Murray and J. L. Arnott, “Toward the simulation of
emotion in synthetic speech: A review of the literature on human vocal emotion,” J. Acoust. Soc. Am., 93, 1097–1108 (1993).

[20] M. Schroeder, “Expressing degree of activation in synthetic speech,” IEEE Trans. Audio Speech Lang. Process., 14, 1128–1136 (2006).

[21] H. Mori, H. Aizawa and H. Kasuya, “Consistency and agreement of paralinguistic information annotation for conversational speech,” J. Acoust. Soc. Jpn. (J), 61, 690–697 (2005) (in Japanese).

[22] http://research.nii.ac.jp/src/en/UUDB.html

[23] T. Nose and T. Kobayashi, “A technique for estimating auditory attention cues and task-dependent high level information,” IEEE Trans. Audio Speech Lang. Process., 17, 1009–1034 (2009).

[24] M. Avanzi, A. L. Dujour and B. Victorri, “A corpus-based learning method for prominence detection in spontaneous speech,” Proc. Speech Prosody 2010, Chicago, 4 pp. (2010).

[25] L. Schillingmann, P. Wagner, C. Munier, B. Wrede and K. Rohlfing, “Using prominence detection to generate acoustic feedback in tutoring scenarios,” Proc. INTERSPEECH 2011, Florence, pp. 3105–3108 (2011).

[26] A. Ortony, G. L. Clore and A. Collins, The Cognitive Structure of Emotions (Cambridge University Press, Cambridge, 1998).

[27] A. Yamamoto, K. Suzuki, K. Cho and Y. Yamashita, “Automatic prosodic labeling of accentuated syllables,” Proc. INTERSPEECH 2006, Aalborg, pp. 87–90 (2001).

[28] J. E. Cahn, “The generation of affect in synthesized speech,” J. Am. Voice I/O Soc., 8, 1–19 (1990).

[29] Y. Kituhara and Y. Tohkura, “Prosodic control to express emotions for man-machine speech interaction,” IEICE Trans. Fundam., E75-A, 155–163 (1992).

[30] I. R. Murray and J. L. Arnott, “Implementation and testing of a system for producing emotion-by-rule in synthetic speech,” Speech Commun., 16, 369–390 (1995).

[31] F. Burkhardt and W. F. Sendilmier, “Verification of acoustical correlates of emotional speech using formant-synthesis,” Proc. ITRW on Speech and Emotion, Newcastle, pp. 151–156 (2000).

[32] E. Zovato, S. Sandri, S. Quazza and L. Badino, “Prosodic analysis of a multi-style corpus in the perspective of emotional speech synthesis,” Proc. INTERSPEECH 2004, Jeju, pp. 1897–1900 (2004).

[33] M. Bulut, S. Narayanan and A. K. Syrdal, “Expressive speech synthesis using a concatenative synthesizer,” Proc. ICSLP 2002, Denver, pp. 1265–1268 (2002).

[34] A. Iida, N. Campbell, F. Higuchi and M. Yasumura, “A corpus-based speech synthesis system with emotion,” Speech Commun., 40, 161–187 (2003).

[35] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza and M. A. Picheny, “The IBM expressive text-to-speech synthesis system for American English,” IEEE Trans. Audio Speech Lang. Process., 14, 1099–1108 (2006).

[36] T. Moriyama, S. Mori and S. Ozawa, “A synthesis method of emotional speech using subspace constraints in prosody,” J. Inf. Process., 50, 119–1191 (2009) (in Japanese).

[37] S. Mori, T. Moriyama and S. Ozawa, “Emotional speech synthesis using subspace constraints in prosody,” Proc. 2006 IEEE Int. Conf. Multimedia and Expo, pp. 1093–1096 (2006).

[38] H. Zen, K. Tokuda and A. W. Black, “Review: Statistical parametric speech synthesis,” Speech Commun., 51, 1039–1064 (2009).

[39] K. Tokuda, “Recent advances in statistical parametric speech synthesis,” J. Acoust. Soc. Jpn. (J), 67, 17–22 (2011) (in Japanese).

[40] K. Hirose, K. Sato, Y. Asano and N. Minematsu, “Synthesis of F0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis,” Speech Commun., 46, 385–404 (2005).

[41] J. Yamagishi, M. Tachibana, M. Masuko and T. Kobayashi, “Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis,” Proc. ICASSP 2004, Montreal, I, pp. 5–8 (2004).

[42] M. Tachibana, J. Yamagishi, T. Masuko and T. Kobayashi, “Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing,” IEICE Trans. Inf. Syst., E88-D, 2484–2491 (2005).

[43] T. Nose, J. Yamagishi, T. Masuko and T. Kobayashi, “A style
control technique for HMM-based expressive speech synthesis,” *IEICE Trans. Inf. Syst.*, E90-D, 1406–1413 (2007).

[58] T. Nose, M. Tachibana and T. Kobayashi, “HMM-based style control for expressive speech synthesis with arbitrary speaker’s voice using model adaptation,” *IEICE Trans. Inf. Syst.*, E92-D, 489–497 (2009).

[59] R. Tsuzuki, H. Zen, K. Tokuda and T. Kitamura, “Constructing emotional speech synthesizer with limited speech database,” *Proc. ICSLP 2004*, Jeju, pp. 1185–1188 (2004).

[60] T. Nose and T. Kobayashi, “A perceptual expressivity modeling technique for speech synthesis based on multiple-regression HSMM,” *Proc. INTERSPEECH 2011*, Florence, pp. 109–112 (2011).

[61] J. Hirschberg, “Pitch accent in context: Predicting intonational prominence from text,” *Artif. Intell.*, 63, 305–340 (1993).

[62] S. Takeda and A. Ichikawa, “Analysis of prominence in spoken Japanese sentences and application to text-to-speech synthesis,” *Speech Commun.*, 14, 171–196 (1994).

[63] W. Gu, K. Hirose and H. Fujisaki, “Modeling the effects of emphasis and question on fundamental frequency contours of cantonese utterances,” *IEEE Trans. Audio Speech Lang. Process.*, 14, 1155–1170 (2006).

[64] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King and D. Jurafsky, “Modeling prominence and emphasis improves unit-selection synthesis,” *Proc. INTERSPEECH 2007*, Antwerp, pp. 1282–1285 (2007).

[65] L. Badino, J. S. Andersson, J. Yamagishi and R. A. J. Clark, “Identification of contrast and its emphatic realization in HMM based speech synthesis,” *Proc. INTERSPEECH 2009*, Brighton, pp. 520–523 (2009).

[66] Y. Yamashita and R. Mizoguchi, “Dialog management for speech output system from concept representation,” *Proc. ICSLP 1992*, Banff, pp. 1435–1438 (1992).

[67] Y. Asano and K. Hirose, “A dialogue processing system for speech response with high adaptability to dialogue topics,” *IEICE Trans. Inf. Syst.*, E76-D, 95–105 (1993).

[68] http://www.apple.com/iphone/features/siri.html

[69] http://www.nttdocomo.co.jp/english/service/func_tool/index.html

Yoichi Yamashita received his B.E., M.E. and Dr.Eng. degrees from Osaka University in 1982, 1984 and 1993, respectively. He has worked for the Institute of Scientific and Industrial Research of Osaka University as a Technical Official, a Research Associate, and an Assistant Professor from 1984 to 1997. In 1997, he joined Ritsumeikan University as an Associate Professor in the College of Science and Engineering. He is currently a Professor in the College of Information Science and Engineering. His research interests include speech understanding, speech synthesis, speech communication, and spoken document processing. He is a member of IEICE, IPSJ, JSAI, ISCA, and IEEE.