Emotional Speech Synthesis by Sensing Affective Information from Text

[1]Mostafa Al Masum Shaikh, [1]Antonio Rui Ferreira Rebordao, [1]Keikichi Hirose and [2]Mitsuru Ishizuka
Department of Information and Communication Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan
[1]{almasum, antonio, hirose}@gavo.t.u-tokyo.ac.jp
[2]ishizuka@i.u-tokyo.ac.jp

Abstract

Speech can express subjective meanings and intents that, in order to be fully understood, rely heavily in its affective perception. Some Text-to-Speech (TTS) systems reveal weaknesses in their emotional expressivity but this situation can be improved by a better parametrization of the acoustic and prosodic parameters. This paper describes an approach for better emotional expressivity in a speech synthesizer. Our technique uses several linguistic resources that can recognize emotions in a text and assigns appropriate parameters to the synthesizer to carry out a suitable speech synthesis. For evaluation purposes we considered the MARY TTS system to readout "happy" and "sad" news. The preliminary perceptual test results are encouraging and human judges, by listening to the synthesized speech obtained with our approach, could perceive "happy" emotions much better than compared to when they listened non-affective synthesized speech.

1. Introduction

Emotions can add some liveliness to speech and this is an acknowledged issue [3, 4, 10] that should be taken into account when we consider Speech Synthesis. Expressive eloquence also contributes to the naturalness of synthesized speech as indicated by many studies like [3, 8, 11, 13, 14]. It is generally accepted that an unified tone, a proper pitch accent and a suitable intensity of speech can help conveying speech subtleties and their intent in a contextually and content-rich manner. Therefore, if a Text-To-Speech (TTS) system can generate human-like speech, then it can convince or appeal to a particular audience more successfully. Thus, in our opinion a TTS system should produce synthesized speech that resembles speech produced by human articulation but contemporary TTS systems tend to produce synthetic speech in a way that sounds unnatural. This is partly due to some deficiencies in the syntactic analysis of the raw input text and to a lack of semantic information, affective clues, and world knowledge. Several perceptual and objective experiments, that have been carried out in [17], show that the present TTS systems are weak in the characterization and expression of emotions. In [17] the authors provided affective and non-affective text to several state-of-the-art TTS systems and analyzed the synthesized speech samples. This study revealed that the pitch accent assignments in the synthesized speech were inappropriate and that their pitches were very similar to the synthesized speech samples produced out of non-affective sentences. The affective texts had obvious affective connotation (e.g., sad/happy) but this emotional content were not present. Therefore, it is inferred that TTS systems usually fail in encoding emotional connotation in synthesized speech.

Some TTS systems accept XML-like mark-up input text pre-marked with intonational information but we have noticed that very few systems make intelligent text preprocessing that can assist the synthesis process. Our research finds the niche at this point. We use commonsense knowledge and emotion recognition techniques to process the text, annotate appropriate pitch accent to words and/or phrases and adjust the prosodic parameters before the synthesis. For example, the studies [3, 8, 10] show that in order to signal "sadness", the Speech Rate (SR) (i.e., syllable/sec) and Pitch Average (PA) should be slightly slower; the Pitch Range (PR) should be slightly narrower; the Intensity of the signal should be lower; and the Pitch Change (PC) should have downward inflections with respect to neutral speech. To signal "happiness" the SR should be faster or slower; the PA should be much higher; the PR should be much wider; the Intensity should be higher; and the PC should have smooth upward inflections. Therefore our approach tries to configure dynamically those parameters by sensing the affective meaning of the input text. Although automatic prosody control in a TTS is not new, most of the previous work gave emphasis at an acoustic level and not in a text-processing level. So our primary contribution lies in this text-processing zone. Extensive linguistic processing is done at this level and appropriate speech parameters are as-

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

signed that can assist the synthesizer in generating emotion-embedded speech.

This paper is organized as follows. Section 2 discusses the research background and concept. Our approach is described in section 3 and section 4 explains the data-set and the experimental results. Section 5 concludes the paper with some prospects for our future work.

2. Background

Tremendous efforts were done in speech synthesis from text and in identifying emotions in human speech but, as far as we know, there is no system that takes the content (e.g., typed text from a speaking-impaired person like Stephen Hawking) and generates automatically the affective values of the content in order to feed a TTS engine. After a careful review of the existing literature it is found that research regarding expressivity in syntactic speech is closely related to concepts like: emotional text-to-speech synthesis; control languages that guide the TTS synthesis process; flexibility in TTS architecture; and emotion recognition from textual data. The following sub-sections briefly discuss these concepts.

2.1. Emotional Speech Synthesis

Previous researches (e.g., [8, 10, 11, 13, 14]) have found that there are several features in natural speech that are associated with emotions. These features consist in different statistical values (e.g., max, mean, standard deviation, etc.) of the fundamental frequency F0, different statistical values of the first three formants (F1, F2, and F2) and their bandwidths (BW1, BW2, and BW3), energy, speaking rate, etc. These features are generally obtained by observing how human voice changes according to different emotions. Several studies have established the fact that when a speaker is in a state of fear, anger or joy, then his speech is typically faster, louder, and enunciated, with strong high-frequency energy. When the speaker is bored or sad, then his speech is typically slower and low-pitched, with very little high-frequency energy. Such a pragmatic knowledge obtained from speech signal processing has inspired various kinds of synthesis methods like: formant synthesis; diphone concatenation; unit selection; and prosody rules based synthesis. In [13, 14] these techniques are described along with their advantages and disadvantages. Moreover the following approaches attempted to incorporate emotional underpinning in the syntactic speech.

2.1.1 Explicit Prosody Control

Explicit prosody models have been formulated based on various sources of information. Usually, rules are based on a relevant set of acoustic parameters reported in studies like [2, 3, 8]. The system "Affect Editor" [3] is an example of an explicit prosody model where the user needs to adjust the affect parameters by hand. Affect Editor takes an acoustic and linguistic description of an utterance and generates synthesizer instructions for a DECtalk3 synthesizer to produce speech with the desired affect. It is generally agreed that F0 level and range, speech tempo, and loudness are important prosodic settings that indicate emotional expressiveness. There are studies [2, 9] that investigated other parameters, for example, the steepness of the F0 contour during rises and falls, the distinction between articulation rate and the number and duration of pauses, or modeling additional phenomenons like voice quality or articulatory precision.

2.1.2 Expressivity based Unit-Selection

This approach deals with unit selection synthesis where targets related with symbolic expressivity (i.e., intended speaking style) are taken into account during the unit selection process along with the acoustic expressive targets that possess specific phone identity, context, position in the sentence, and so on. Acoustic expressive targets usually rely on acoustical models of expressive styles to identify units that could be suitable for the targeted expressive style indicated by the symbol. This approach was used in some works mentioned in [7].

2.1.3 Unit Selection & Signal Modification

This explicit modeling technique is used in diphone synthesis, and usually avoided in unit selection because of the deteriorating effects on the overall speech quality. But, the authors of [24] applied this technique in unit selection by modifying the pitch and the duration according to prosodic rules related with emotion. As for diphone synthesis, this approach has the disadvantage of not being able to modify voice quality, and of creating audible distortions of larger modifications. To overcome the voice quality drawback, the authors of [6] have proposed a method for modifying the glottal source spectrum, described by the parameters glottal formant and spectral tilt. They decompose the speech signal into a periodic and an aperiodic part, and recombine these after modifying them separately.

2.1.4 HMM based Parametric Synthesis

This new synthesis technology proposed in [23] appeared first in the Blizzard speech synthesis competition in 2005. The core behind the technology is that context-dependent HMMs are trained on a speech database; the spectrum, F0, and duration are modeled separately. The context-dependent models are organized in a decision tree; at runtime, for a given "target" context to be realized, the tree yields the appropriate HMM state sequence corresponding to that context, describing mean and standard deviation of the acoustic features. A vocoding technique is used to generate an audio signal from the acoustic features, resulting in a very intelligible speech output. This approach requires style-specific recordings as in unit selection but offers greater flexibility.

2.1.5 Non-Verbal Vocalizations

Campbell [5] showed that a large proportion of everyday vocalizations are nonverbal: laughs, ”grunts” and other small sounds feed the communication but are not necessarily described as text with prosody. This indicates that expressive conversational speech does not always result from applying suitable prosody modifications. In [4] the author produced a conversational speech synthesizer that uses a huge database of everyday speech as a unit selection database. With adequate annotation of speech units and careful manual selection, this system can produce conversational speech of unprecedented naturalness. The authors of [15] also investigated the suitability of various affect bursts (i.e., short emotional interjections) in a brief conversation.

2.2. XML-based Markup Languages for TTS

XML-based markup languages can be used to add information to a text in order to improve the way it is spoken. These languages are independent of any TTS system and the synthesis processors are assumed to parse this kind of input and translate it into a system-internal data representation format which in most cases is not XML-based.

2.2.1 W3C SSML

The Speech Synthesis Markup Language (SSML) [20] developed by the Voice Browser Working Group is one of the standards for providing a rich, XML-based markup language that assists the generation of synthetic speech for the Web and/or for other applications that allow Web access based on a speech interface. The essential role of this markup language is to provide a standard way to control speech aspects like pronunciation, volume, pitch, rate, etc., across different synthesis-capable platforms. This XML-based language supports tags like, ”break”, ”emphasis”, ”phoneme”, ”prosody”, ”say-as”, ”voice”, etc. that are used as input by some synthesis processors and outputted as speech samples.

2.2.2 SAPI

The Speech Application Programming Interface (SAPI) was developed by Microsoft for speech recognition and speech synthesis within Windows applications. Several SAPI versions have been released as part of a Speech SDK or Windows Operating System. Applications that use SAPI include Microsoft Office, Microsoft Agent and Microsoft Speech Server. All versions have been designed in a way that external engines can work with SAPI (as long they conform to the defined interfaces). SAPI versions 1 through 4 are similar to each other but SAPI 5 was released in 2000 and has a completely new interface. Since then several subversions of this API have been released [22].

2.2.3 MARY XML

The MARY system [1, 16] uses an internal XML-based representation language called MaryXML that provides a powerful method for controlling the behavior of a TTS system. The syntax of a MaryXML document reflects the information required by the modules in the TTS system, such as sentence boundaries and global prosodic settings as used in SSML specification. For example, the following MaryXML example can be used to control prosody, accent and boundary for the articulation of the phrase ”Look at me!”.

```
<?xml version=”1.0” encoding=”UTF-8”?>
<maryxml
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xmlns=”http://mary.dfki.de/2002/MaryXML”/version=”0.4”
xml:lang=”en”>
<prosody rate=”+30%” pitch=”+50%” range=”-5%”
volume=”loud”><t accent=”L+H*”>please</t><t accent
=”L-L%”>Look at me!</t><boundary duration=”100”/>
</prosody>
</maryxml>
```

2.3. Emotion Sensing From Text

This research addresses the aspect of subjective opinion and it includes the identification of different emotive dimensions and text classification by emotion affinity (e.g. happy, sad, anger, etc). It can be argued that analyzing attitude and affect in texts is an ”NLP”-complete problem and its interpretation depends on audience, context, and world knowledge. The approaches for assessing affective information from text are based in one or in a combination of the following techniques: keyword spotting; lexical affinity; statistical methods; a dictionary of affective concepts and lexicon; commonsense knowledgebase; fuzzy logic; knowledge-base from facial expression; machine learning; domain specific classification; and contextual valence assignment. The researches [12, 18, 21] reported these techniques extensively. Shaikh et al. [18] have implemented contextual valence assignment technique and achieved tremendous results in emotion recognition from text.

2.4. Mary TTS System

The MARY system [16] is a TTS client-server application written in Java and created at DFKI. MaryXML serves as the configuration input language of the system and this system became a very flexible toolkit for speech synthesis research. This is the main reason why we have chosen it. This system allows dynamic creation of MaryXML with appropriate prosodic and suitable accent properties and all intermediate processing results can be accessed for purposes of debugging and analysis.

3. Our Approach

The vocal intonation of how something is said elucidates two aspects: cues emphasizing the content in the message that is most important, and cues arising from the speaker’s affective state. From linguistic standpoints our approach targets both components. The first step is to assess the affective content of the input text, then prosodic parameters related with that affective content (”positive” and ”negative” emotions) are setup according to the findings reported in [3, 8, 10, 11, 13] and the last step is a proper assignment of phrasal tones (e.g., L-, H-, etc.) and pitch accents (e.g., H*, L+H*, etc.) by ToBI notation [19]. The goal is to produce MaryXML that can generate synthetic speech with a suitable emotional expressivity.

3.1. System Architecture

A pipeline architecture with the following steps is followed: Affect Sensing from Text; Prosodic Parameterisation; Pitch Accent Annotator, and Dynamic MaryXML creation. These steps are briefly described as following.

3.1.1 Affect Sensing from Text

We have used the output of the system SenseNet developed by Shaikh et al. [18]. SenseNet can perform sentence level affect sensing by assessing the contextual valence of the words using rules and prior valence values of the words. It outputs a numerical value ranging from -15 to +15 flagged as the ”sentence-valence” for each input sentence. As examples, SenseNet outputs -8.96 and +10.47 for the inputs: *”A terrorist attack on Britain with chemical, nuclear or biological weapons is now more realistic because of the increasing theft of materials used to make a dirty bomb, the Government warned yesterday”* and *”With help from IBM, Cisco, Philips and other companies, the city’s infrastructure is becoming ultra energy-efficient, attracting global attention”*, respectively. The output value indicates a numerical measure of negative or positive sentiments carried by the sentence. In this case we consider negative or positive score for a given sentence as ”sad” or ”happy”, respectively. SenseNet outputs the valence values of the events mentioned in the input sentence. The concept of an event is a triplet comprising of subject-verb-object-object. For example, for the negative sentence in the above example four triplets are outputted as following: terrorist attack is now more realistic, the increasing theft of materials used, make bomb, government warned yesterday. The first triplet has attributes like: ”Britain”, ”chemical, nuclear or biological weapons” and each of them is given a numerical score. The system applies several computational linguistic rules to assign a contextual valence for the triplet. Thus triplet 1 gets a negative score due to the indication of a negative action along with other entities. It also outputs several things like, whether the event is ”praiseworthy” or ”blameworthy” or ”common” or ”uncommon” or if the object of an event is either ”attractive” or ”not-attractive”. The accuracy of SenseNet to assess sentence-level negative/positive sentiments is 91% and the classification accuracy of eight emotion types is 82% in an experimental study [18].

3.1.2 Prosodic Parameterisation

After the input text has been processed as mentioned above, we obtain its affective assessment like: the overall emotion carried by the text; the positive or negative connotations of the events mentioned in the text; the attributes (e.g., location, time, etc.) of the events; the quality of the events (i.e., praiseworthy/ blameworthy, common/uncommon); and the quality of the targeted object related with the event. After this, several speech parameters are set that match the overall affective connotation of the text (compared to neutral emotion). The process follows the findings of [3, 10, 11, 14]. For example, if the text would have to meant ”happy”, then the overall speech rate is set faster, the pitch average is set higher, the pitch range is set much wider, the intensity is higher, and the pitch changes are defined as smooth upward. MaxyXML offers a rich set of prosodic attributes that can be attributed.

3.1.3 Pitch Accent Annotator

Then some rules are applied to assign suitable phrasal tones and pitch accents to be processed by the synthesizer during the synthesis. Such markups are incorporated by MaryXML’s tag support for accent control. Phrasal tones are assigned at every intermediate or intonation phrase. Four types of phrasal tones, L-L%, L-H%, H-H%, and H-L% are considered. The rules to annotate phrasal tones are:

- Tones are assigned by considering verb-phrase, noun-phrase and object-phrase at Triplet level. A Triplet basically has two parts, the event and event’s attribute. Both event and event attributes may have affective values. Based on their affective values tones are annotated.
- If an event shows negative affect and is associated with a positive actor (e.g., car exploded), then H-L%, else L-L% (e.g., the terrorist killed). But, if an event shows positive affect associated with positive actor and action (i.e., infrastructure is becoming ultra energy-efficient), then H-H%, else if both actor and action are negative (e.g., suicide blast killed five outlaws), then L-H% is assigned.
- If an event shows negative affect associated with a positive action and negative object (e.g., sending suicide-bomber), then H-L% else, if an event has negative action with a positive object (e.g., killing people), then L-L% is assigned.
- If an event’s attribute has a positive adjective with a positive entity (e.g., now more realistic), then H-H%. If negative adjective with either positive or negative entity (e.g., alone in the apartment, terrible murder), then L-L%. If positive adjective with negative entity (e.g., popular crime zone), then H-L% is assigned.

Pitch accent tones are marked at every accented syllable. The system annotates for peak accent (H*), low accent (L*), scooped accent (L*+H) and rising peak accent (L+H*) on word level considering whether the word represents a verb or object or attribute cue of an event. In this case some of the rules are:

- If a verb word has negative value associated with an event having certain values of blameworthy and uncommon variables, then L+H is assigned;
- If a verb word has positive value associated with an event having certain values of praiseworthy and uncommon variables, then L+H is assigned;
- If a verb has positive or negative meaning but the event doesn't have certain values for either of those two variables, then H* and L* are assigned respectively;
- If the word is an attribute cue (e.g., near, on) that complements the description of location, time, etc. of an event, then H* assigned to emphasize it as vital information that needs to be spoken.

Similar rules like verb words are applied to nouns considering the value of ”attractive” variables. Details are not given due to space limitation.

3.2. Interconnecting with MARY TTS

At present the MARY TTS system has the following natural language components namely: Tokenize; Preprocessing; and Tagger & Chunker. These can process an input given in MaryXML format. Our system, at present, has nothing to deal with these components rather than creating MaryXML formatted input from a given plain text. In future we plan to add a pre-processing component to the MARY system that would do emotion recognition from the plain text and automatically generates MaryXML in accordance to the recognized emotion.

4. Test and Evaluation

A preliminary perceptual test was conducted in order to assess the validity of the proposed approach.

4.1. Data Set

The data set used in the perceptual test consists of 40 synthesized speech audio samples that were produced by the MARY TTS out of 20 online news collected from RSS feeds of several sources of online news like BBC News, etc. For each news text we created 2 versions of synthesized speech samples using the MARY TTS system. One is the output obtained from the plain text input and the other is produced by inputting dynamic MaryXML pre-marked with intonational information created by our approach. Both cases used the voice Mbrola-us2 version 3.5.0 and the length of each synthesized speech audio sample is 7 seconds on average. The Mbrola voices are the unique choice for our approach because they take the markup prosody into account.

4.2. Experiments Procedure

The survey was conducted online at http://research.rebordao.net/emonews/ and a total of 30 people participated. The subjects had to listen to the synthesized speech audio samples produced from the plain text and from the dynamic MaryXML input and assessed if they could perceive any emotion, or not. If an emotion would be perceived, it would be asked them to classify it as either ”happy” or ”sad”.

4.3. Results and Discussion

We have two systems, the plain text input system (S1) and the dynamic MaryXML input system (S2). We considered the scores obtained from the web-survey for which, either one or both systems, received an emotion perception score. From Chi-Square test it is evident that the evaluation scores of the systems are statistically significant ($P<0.001$) in terms of emotional expressivity. Figure 1 shows that the systems performed almost similarly (i.e., accuracy 52.2%, 47.3% for S1 and S2 respectively) for conveying ”Sad” emotion. But Figure 2 shows that S2 performed significantly better than S1 (i.e., accuracy 6.0%, 67.4% for S1 and S2 respectively) to convey ”Happy” emotions. The case in Figure 1 happened due to the tendency of S1 to produce synthesized speech with intonational information related to negative emotions and that is why the subjects perceived the output of S1 as ”Sad” most of the time. The results are encouraging in two manners, firstly S1 is very weak to convey

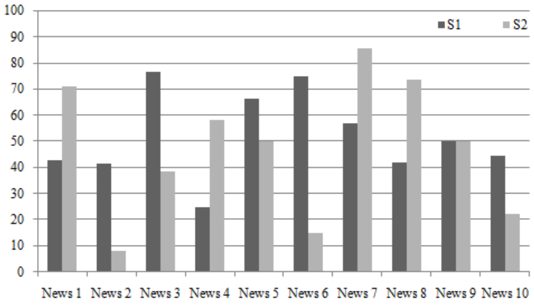

Figure 1. The recognition rates of S1 and S2 for 10 sad news.

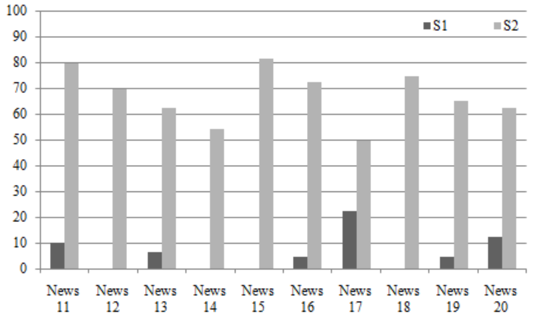

Figure 2. The recognition rates of S1 and S2 for 10 positive news.

positive emotions (e.g., "happiness"), so our approach can solve this problem and secondly, S1 has a tendency to express negative emotions (e.g., "sadness") and our approach can be applied to incorporate different levels of negativism within the phrases of a synthesized sentence.

5. Conclusion

There are numerous research works and techniques to incorporate expressiveness in synthesized speech and this can be achieved by creating speech that conveys suitable emotions. In this paper we have found that the synthesized speech samples produced by a well-known TTS (i.e., MARY TTS) from plain text are not affectively expressive. However, this problem can be solved by pre-processing the input in two manners, first by recognizing the emotion conveyed through the plain text and then controlling the synthesis process by assigning appropriate prosodic parameters that suit the detected emotions. Thus the output of our approach is an enhanced XML-based (i.e., dynamic MaryXML) interpretation of the simple input text that is given to the TTS system (i.e., MARY TTS) to process. Some perceptual tests were performed using synthesized speech samples produced with our approach and the evaluation supports that these speech samples are more affectively expressive than the speech samples synthesized from the plain text. As future work we plan to build a tool combining all the resources discussed in our approach and add it to the MARY TTS system. It would allow a speech impaired person to type text and generate synthesized speech conveying appropriate emotions. Although this paper dealt just with two emotions, in the near future we plan to consider more types of emotions.

References

[1] Mary TTS System. [Online; accessed 22-March-2009].

[2] F. Burkhardt and W. F. Sendlmeier. Verification of acoustical correlates of emotional speech using formant synthesis. In *In Proceedings of the ISCA Workshop on Speech and Emotion*, pages 151–156, Northern Ireland, 2000.

[3] J. E. Cahn. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19, 1990.

[4] N. Campbell. Developments in corpus-based speech synthesis: Approaching natural conversational speech. *IEICE - Trans. Inf. Syst.*, E88-D(3):376–383, 2005.

[5] N. Campbell. Approaches to conversational speech rhythm: Speech activity in two-person telephone dialogues. 2008.

[6] C. d'Alessandro and B. Doval. Voice quality modification for emotional speech synthesis. In *In Eurospeech 2003*, pages 1653–1656, Bonn, Germany, 2003.

[7] R. Fernandez and B. Ramabhadran. Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, pages 34–39, Bonn, Germany, 2007.

[8] D. Morrison, R. Wang, and L. C. de Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech Commun.*, 49(2):98–112, 2007.

[9] S. J. L. Mozziconacci and D. J. Hermes. Role of intonation patterns in conveying emotion in speech. In *in Proceedings of ICPhS*, pages 2001–2004, 1999.

[10] I. R. Murray and J. L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[11] P.-Y. Oudeyer. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59:157–183, 2002.

[12] B. Pang and L. Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[13] M. Schroder. Approaches to emotional expressivity in synthetic speech. In K. Izdebski, editor, *Emotions in the Human Voice: Culture and Perception*, pages 307–321. Plural Publishing, 2008.

[14] M. Schroder. *Affective Information Processing*, chapter Expressive Speech Synthesis: Past, Present and Possible Futures, pages 111–126. Springer, 2009.

[15] M. Schroder, D. K. J. Heylen, and I. Poggi. Perception of non-verbal emotional listener feedback. In R. Hoffmann and H. Mixdorff, editors, *Speech Prosody 2006, Dresden*, volume 40 of *Studientexte zur Sprachkommunication*, pages 43–46, Dresden, 2006. TUDpress.

[16] M. Schroder and J. Trouvain. The german text-to-speech synthesis system mary: A tool for research, development and teaching: Special issue on speech synthesis: Part ii. *International Journal of Speech Technology*, 6:365–377(13), October 2003.

[17] M. A.-M. Shaikh, M. K. I. Molla, and K. Hirose. Assigning suitable phrasal tones and pitch accents by sensing affective information from text to synthesize human-like speech. In *Proceedings of InterSpeech*, pages 326–329, Brisbane, 2008.

[18] M. A. M. Shaikh, H. Prendinger, and M. Ishizuka. Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. *Applied Artificial Intelligence*, 22(6):558–601, 2008.

[19] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirsberg. ToBI: A standard for labeling english prosody. In *Proc. ICSLP*, pages 867–870, 1992.

[20] W3C. Speech synthesis markup language (ssml) version 1.0. [Online; accessed 25-March-2009].

[21] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):0, 2005.

[22] Wikipedia. Speech application programming interface, 2009. [Online; accessed 26-June-2009].

[23] H. Zen and T. Toda. An overview of nitech hmm-based speech synthesis system for the blizzard challenge 2005. In *In Proceedings of InterSpeech 2005*, pages 93–96, Lisbon, Portugal, 2005.

[24] E. Zovato, A. Pacchiotti, S. Quazza, and S. S. Towards emotional speech synthesis: A rule based approach., 2004.