The Speect text-to-speech system entry for the Blizzard Challenge 2013

Johannes A. Louw, Georg I. Schlünz, Willem van der Walt, Febe de Wet, Laurette Pretorius

Human Language Technologies Research Group
Meraka Institute, CSIR, Pretoria, South Africa
jalouw@csir.co.za, gschlunz@csir.co.za

Abstract

This paper describes the Speect text-to-speech system entry for the Blizzard Challenge 2013. The techniques applied for the tasks of the challenge are described as well as the implementation details for the alignment of the audio books and the text-to-speech system modules. The results of the evaluations are given and discussed.

Index Terms: speech synthesis, multilingual, open source, Blizzard Challenge 2013, audio books.

1. Introduction

This paper presents our second entry into the Blizzard Challenge [1], where different speech synthesis techniques are subjectively evaluated and can be directly compared due to the use of a common corpus of speech data.

The training data given to the participants consisted of (A) a set of unsegmented audio ($\pm$ 300 hours) from audio books read by one speaker and (B) a set of 9741 ($\pm$ 19 hours) segmented waveforms, extracted from different audio books (”Black Beauty” and ”Mansfield Park”), read by the same speaker. The segmented audio was accompanied by a text file from which the text of these utterances could be reconstructed. Our group only participated in the English language tasks (EH1 and EH2) which were to build a voice from the two supplied data sets or selections thereof.

All the audio book alignments, and utterance and label generation modules were implemented in-house and only the supplied recordings and text annotations (for the segmented data) were used, thus excluding the other labels provided and limiting manual intervention.

The remainder of the paper is organised as follows, Section 2 gives an overview of our text-to-speech (TTS) system, Section 3 describes our methods for building the voices for the tasks, Section 4 presents the results followed by a discussion and conclusion in Section 5.

2. System Overview

2.1. The Speect TTS system

The architecture of Speect, our TTS system has been reported on in detail in previous publications [2, 3] and apart from the basic modules described in [3] we have added or changed the modules described in the following sections.

2.2. Natural Language Processing

2.2.1. Lexicon and Stress

The Carnegie Mellon University Pronunciation Dictionary (“http://www.speech.cs.cmu.edu/cgi-bin/cmudict”) was used for lexical look-up as well as stress assignment.

2.2.2. Pronunciation prediction

A set of letter-to-sound rewrite rules were trained from “cmudict” using the Default&Refine algorithm [4].

2.2.3. Part-of-speech tagging and Phrasing

The part-of-speech (POS) tagging and phrasing was done using the Stanford CoreNLP tools [5], including the the named entity recognizer (NER) [6], and the coreference resolution system [7].

Recent literature on prosodic modelling [8, 9, 10, 11, 12, 13] point toward the influence of higher linguistic processes, namely discourse, information structure and affect. In particular, [11] has applied the cognitive theory of [14], also known as the *OCC model*, to affect detection from text. However, they have not been successful at modelling diverse emotions in synthesised speech [12]. Our Blizzard 2013 entry reimplements the OCC model in an attempt to remedy this.

2.2.4. The OCC Model

Simplistically, the OCC model appraises human emotions from valenced reactions to three aspects of the environment:

1. The *consequence* of an event—whether it is desirable or undesirable with respect to one’s *goals*.

2. The *action* of the agent responsible for the event—whether it is praiseworthy or blameworthy with respect to one’s *standards*.

3. The *aspect* of an object—whether it is appealing or unappealing with respect to one's *attitudes* (inter alia tastes).

The goals, standards and attitudes of a person are the cognitive antecedents that determine whether his valenced reaction to the environment is positive or negative. A particular emotion is the consequent of the appraisal process, as the person focuses on either the consequence, action or aspect, respectively.

The OCC model neatly defines the concepts necessary for the eliciting conditions of emotional appraisal: on the one hand the environmental factors of events, agents and objects, and on the other the cognitive antecedents of goals, standards and attitudes. The former group can be inferred from text in a straightforward manner using shallow semantic parsing that identifies the predicate, or action (typically the verb), and assigns roles to the arguments of the predicate. These are predominantly an AGENT role to the entity who performs the action, and a PATIENT role to the entity who undergoes the action. Hence, semantic predicates map to OCC events and semantic AGENTs and PATIENTs to OCC agents or objects. The cognitive antecedents are much harder to infer from text.

2.2.5. Semaffect

It is necessary to rethink the semantically-complex high-level concepts of goals, standards and attitudes in order to come to a tractable solution for the eliciting conditions of the OCC model. Semaffect is a semantically-simple low-level model of affect that aggregates the OCC goals, standards and attitudes into a single sense, or *judgment*, of right and wrong (or good and bad) according to the (subjective) belief system of the person.

Informally, Semaffect appraises an emotion from how one reacts to a good/bad person doing a good/bad deed to another good/bad person. Formally, the model appraises a given event in terms of the good (1) and bad (0) valences of its semantic AGENT (**A**), verb predicate (*v*) and PATIENT (**P**). It is important to note that Semaffect defines an emotion *anonymously* based on the composition of the underlying semantic variables **A**, *v* and **P**, and not from a predefined surface set with particular definitions. Semaffect will model the subjective affective responses of a person accurately as long as the individual remains consistent in his belief system (for example, good always deserves good and bad always deserves bad, et cetera). It does not promise any consistency across different persons, that is objective or absolute affective responses. The number of possible affective states produced by Semaffect is $2^3 = 8$, as illustrated in Table 1. The discourse context for the examples in the table is the following:

```
Policemen are good.
Criminals are bad.
To save someone is good.
To kill someone is bad.        (1)
```

The implementation of Semaffect for discourse text involves certain key design decisions (inter alia assumptions) to put the theory of a person's judgment of right and wrong into practice successfully:

- Right and wrong, good and bad are represented by the boolean values of true (1) and false (0).
- The good or bad valence of a discourse *entity* represented by a noun phrase defaults to the entry of (the lemma of) the head noun in the SentiWordNet lexicon. SentiWordNet [15] assigns a positive or negative sentiment score (true or false) to each WordNet [16] entry. If no entry is available, a good valence is assigned.
- The entity valence may be altered by the SentiWordNet valences of (the lemmas of) modifiers to the head noun (such as adjectives) or negated by negators (such as `not`). Modification happens in a "once bad, always bad" fashion: once a bad valence occurs in the modifier-head noun chain, the entity valence becomes bad. Logically, this is by boolean conjunction (AND). Negation is applied straightforwardly after modification by boolean negation (NOT).
- The good or bad valence of a discourse *action* represented by a verb phrase defaults to the SentiWordNet entry of (the lemma of) the head verb. If no entry is available, a good valence is assigned.
- The action valence may also be altered by the SentiWordNet valences of (the lemmas of) modifiers to the head verb (such as adverbs) or negated by negators (such as `not`). Modification and negation follow the same principles as their entity counterparts.
- As the discourse progresses, the entities and actions can be reassigned valences when they appear in assertive statements as the subjects of copular verbs (for example `to be`). The copula (SentiWordNet entry modified and negated) determines the new valence.
- Semaffect operates on a *clause* level. The narrative is divided into sentences, and the sentences into clauses. Each clause is semantically parsed into a verb predicate with an AGENT and a PATIENT.

Table 1: Possible combinations of valenced semantic states

A	*v*	P	Gloss	Example
0	0	0	bad A doing bad deed to bad P	*The criminal kills another criminal.*
0	0	1	bad A doing bad deed to good P	*The criminal kills the policeman.*
0	1	0	bad A doing good deed to bad P	*The criminal saves another criminal.*
0	1	1	bad A doing good deed to good P	*The criminal saves the policeman.*
1	0	0	good A doing bad deed to bad P	*The policeman kills the criminal.*
1	0	1	good A doing bad deed to good P	*The policeman kills another policeman.*
1	1	0	good A doing good deed to bad P	*The policeman saves the criminal.*
1	1	1	good A doing good deed to good P	*The policeman saves another policeman.*

Valences are calculated (an absent AGENT or PATIENT receives a good valence) and the affective state is appraised.

2.3. Digital Signal Processing

An utterance processor plug-in was written to provide an interface for Speect to the *hts_engine API* (version 1.05) [17]. The vocoder is based on the standard *hts_engine API* vocoder, but has been modified to include mixed excitation.

3. Voice Building

The tasks for the English language data were to build a voice from each of the two data sets supplied (or selections thereof), and are described in more detail below.

3.1. Audio Books

3.1.1. Text

The text corresponding to the audio book training data was not provided and part of the challenge was to locate these texts. The following texts could not be found from free and open resources and their audio was also excluded from further processing:

"On a Flying Fish" by David Applefield, "The Trumpet of the Swan" by E. B. White, "Heidi" by Joanna Spyri, "The White Cat of Drumgonniol" by J Sheridan Le Fanu, "Gentlemen Prefer Blondes" by Anita Loos, "Chéri" by Colette, and "The Gospel According to Condo Don" by Fred Dungan.

The following books were found from sources other than Project Gutenberg (`http://www.gutenberg.org/`):

"Roman Fever" by Edith Wharton (`http://classiclit.about.com/library/bl-etexts/ewharton/bl-ewhar-roman.htm`) and "Scandal" by Willa Cather (`http://www.online-literature.com/willa-cather/2118/`).

"The Facts in the Case" by M. Valdemar was found in "The Works of Edgar Allen Poe". "Bernice bobs her hair" by F. Scott Fitzgerald was found in "Flappers and Philosophers".

3.1.2. Text processing

As many of the books were sourced from Project Gutenberg, a script was written to remove the front and back texts, present in all Gutenberg texts. Where possible, the texts were split into chapters using custom scripts. There were however some books which could not reliably be processed using automatic means and had to be processed manually.

Some of the collections of short stories were slightly different versions from the provided audio, causing a lot of manual manipulation. For some books, there were more audio than text e.g. the audio version included a preface and the text not. In some books the opposite was true, e.g. the text contained a table of contents and the audio did not.

In the case of the "King James Bible", the individual books were retrieved from Project Gutenberg as it was easier to process them into chapters using scripts. The audio however did not contain the verse numbers, so that was removed. In some cases, the audio contained more than one chapter, so the text had to be adjusted accordingly.

3.2. Alignment and HMM training

For each book, the chapter-level text was processed by the Speect frontend and Stanford CoreNLP (`http://nlp.stanford.edu/software/corenlp.shtml`) to detect sentence and clause boundaries, parse the semantic AGENTs, predicates and PATIENTs and calculate the appropriate Semaffect affective states. This information accounts for the linguistic half of the aligned data.

For the acoustic half, the Hidden Markov Model Toolkit (HTK) [18] was used in the forced alignment of the audio to the phonetic transcriptions of each book in multiple phases. In the first phase, the chapter-level audio was aligned to the chapter-level transcriptions using North American English triphone acoustic models trained on the English Broadcast News Speech corpus [19]. In the second phase, these chapter-level alignments were split at the utterance-level, quality controlled and then

used to train speaker-specific triphone acoustic models. In the third and final phase, the chapter-level audio was realigned using the speaker-specific models and, again, split into utterances and quality controlled. Hence, per utterance in the original text and speech of each audiobook, phonetic alignment information is captured along with the linguistic information, ready for input to the HMM-Based Speech Synthesis System (HTS) [20] voice training procedure.

The quality control employed the phone-based dynamic programming (PDP) technique of [21] to score the phonetic alignments. Basically, it computes the condence score of an utterance as the lowest dynamic-programming cost when aligning the freely decoded phone string to the forced alignment of the provided transcription. In particular, [21] specifies the following steps, given an audio and text segment:

1. Free recognition is performed on the audio segment using a phone-loop grammar in order to produce an *observed string*.
2. A dictionary lookup, or an ASR alignment if the target phone string is a segment within a larger utterance, produces a *reference string*.
3. A standard dynamic programming algorithm with a pre-calculated scoring matrix is used to align the observed and reference string with each other. The scoring matrix specifies the cost associated with a specic substitution between a phone in the reference string and the observed string.
4. The resulting score obtained from the best dynamic programming path is divided by the number of phones in the alignment, which may be longer than either of the strings individually.
5. This score is normalised by subtracting the optimal score that can be obtained for the given reference string.

3.2.1. Filtering

In addition to performing the PDP scoring, the quality control in the third phase also filters out all utterances that do not adhere to a strict semantic AGENT-verb-PATIENT structure in their constituent clauses, in an effort to accommodate Semaffect. This brings the total number of hours of training data down from $\pm$ 300 hours to 30 hours for task EH1 (the unsegmented audio voice), while all available data was used for task EH2.

4. Results

Our designated system identification letter is “H”; system “A” is natural speech, system “B” is a Festival unit-selection voice benchmark, system “C” is an HTS benchmark while the other systems are other participants.

We are just reporting on the overall impression, naturalness, similarity to original speaker and word error rate in Figures [1,2,3,4,5,6,7,8] for each of the two voices (EH1 and EH2) for the sake of brevity. The full results will be published in the workshop summary. Note that all the figures give the results from all the participants in the evaluations.

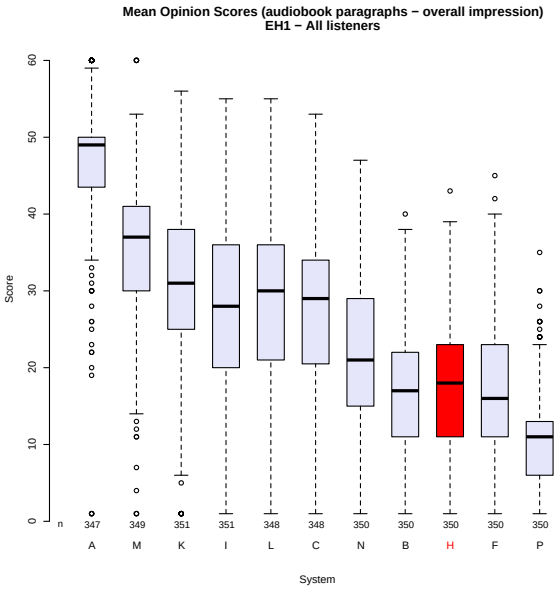

Figure 1: *Task EH1: Overall impression.*

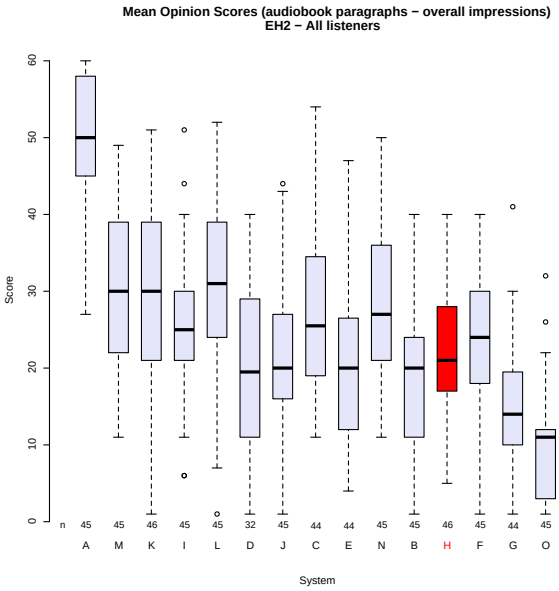

Figure 2: *Task EH2: Overall impression.*

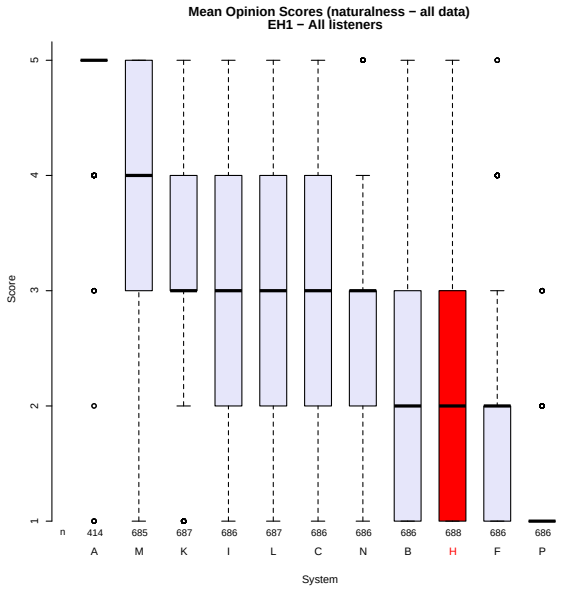

Figure 3: *Task EH1: Naturalness.*

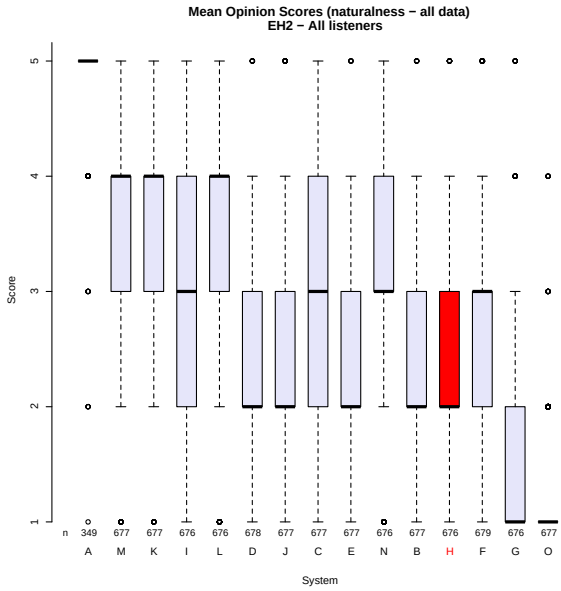

Figure 4: *Task EH2: Naturalness.*

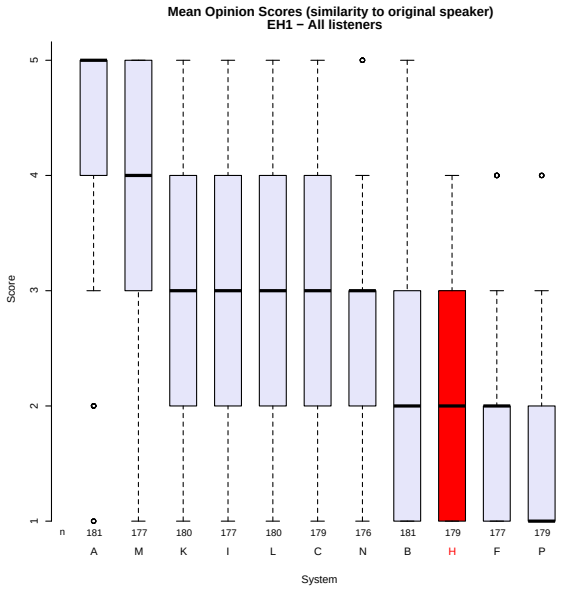

Figure 5: *Task EH1: Similarity to original speaker.*

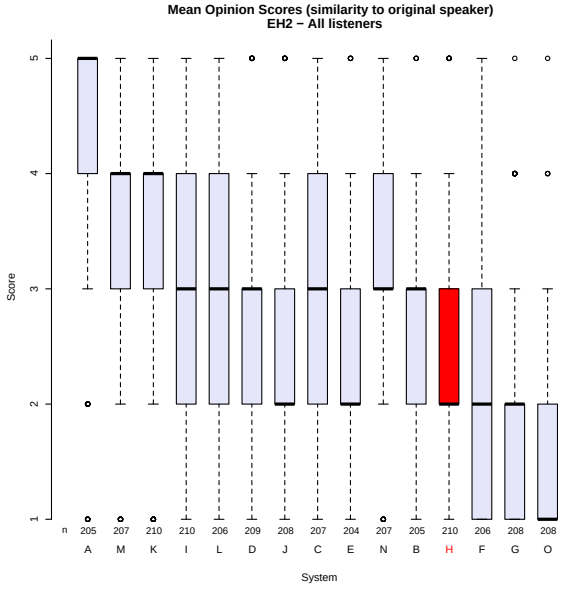

Figure 6: *Task EH2: Similarity to original speaker.*

5. Discussion and conclusion

The general trend from the results are that our system suffers from a lack of naturalness and similarity to the original speaker in comparison to the other systems, but compares favourably in intelligibility (word error rate). This was not unexpected as our system employs only a simple mixed excitation vocoder and not advanced spectral representations (such as STRAIGHT [22]). Our own efforts into an HMM based harmonic plus noise model (HNM) vocoder implementation was not completed in time for the challenge. Our implementation of the OCC model (section 2.2.4) did not deliver the desired results in the realisation of emotion, but we have improved it since the challenge and hope to report better results in future publications.

Our system compared better in task EH2 than task EH1 (with regards to naturalness scores), which might be as a result of our method of filtering the audio data of task EH1 as described in section 3.2.1. The exact reason will need to be investigated.

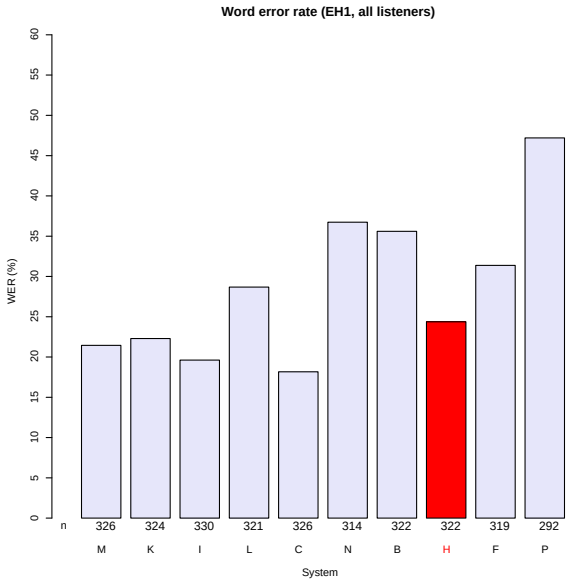

Figure 7: *Task EH1: Word error rate.*

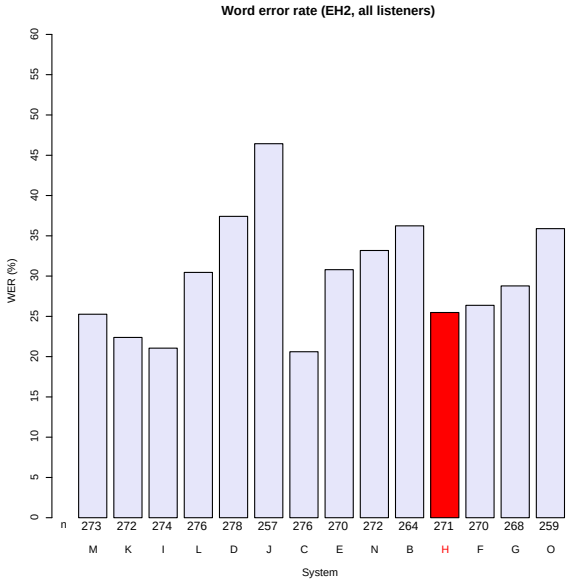

Figure 8: *Task EH2: Word error rate.*

6. References

[1] A. W. Black and K. Tokuda, "The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets," in *Interspeech*, Lisbon, Portugal, 2005.

[2] J. A. Louw, "Speect: a multilingual text-to-speech system," in *Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa*, Cape Town, South Africa, 27-28 November, 2008.

[3] J. A. Louw, D. van Niekerk, and G. Schlunz, "Introducing the Speect speech synthesis platform," in *Blizzard Challenge 2010 Workshop*, Kyoto, Japan, September 25, 2010.

[4] M. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[5] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of HLT-NAACL 2003*, 2003, pp. 252–259.

[6] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," in *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, 2005.

[7] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task," in *Proceedings of the CoNLL-2011 Shared Task*, 2011.

[8] A. Kratzer and E. Selkirk, "Phase theory and prosodic spellout: The case of verbs," *The Linguistic Review*, vol. 24, pp. 93–135, 2007.

[9] M. Steedman, "Information-structural semantics for english intonation," in *Topic and focus: cross-linguistic perspectives on meaning and intonation*, C. Lee, M. Gordon, and D. Büring, Eds. Springer, 2007, pp. 245–264.

[10] C. Féry and S. Ishihara, "How focus and givenness shape prosody," in *Information Structure from Different Perspectives*, M. Zimmermann and C. Féry, Eds. Oxford University Press, 2009, pp. 36–63.

[11] M. Shaikh, H. Prendinger, and M. Ishizuka, *Affective Information Processing*. Springer Science+Business Media LLC, 2009, ch. A Linguistic Interpretation of the OCC Emotion Model for Affect Sensing from Text, pp. 45–73.

[12] M. Shaikh, A. Rebordao, and K. Hirose, "Improving tts synthesis for emotional expressivity by a prosodic parameterization of affect based on linguistic analysis," in *Proceedings of the 5th International Conference on Speech Prosody*, Chicago, USA, 2010.

[13] C. Tseng, "Beyond sentence prosody," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, 2010, pp. 20–29.

[14] A. Ortony, G. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[15] A. Esuli and F. Sebastiani, "SentiWordNet: a publicly available lexical resource for opinion mining," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006, pp. 417–422.

[16] C. Fellbaum, Ed., *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press, 1999.

[17] K. Tokuda, S. Sako, H. Zen, K. Oura, K. Nakamura, and K. Saino, "The hts_engine API version 1.05," http://hts-engine.sourceforge.net/.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department, 2005.

[19] D. Graff, J. Garofolo, J. Fiscus, W. Fisher, and D. Pallett, "1996 english broadcast news speech," Linguistic Data Consortium, Philadelphia, 1997.

[20] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proceedings of ISCA SSW6*, 2007, pp. 294–299.

[21] M. H. Davel, C. J. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *Proceedings of the Third International Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU'12)*, 2012.

[22] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.