A Discourse Model of Affect for Text-to-Speech Synthesis

Georg I. Schlünz*†
*Human Language Technology Research Group
CSIR Meraka Institute, Pretoria, South Africa
gschlunz@csir.co.za

Etienne Barnard†
†Multilingual Speech Technologies Group
North-West University, Vanderbijlpark, South Africa
etienne.barnard@nwu.ac.za

***Abstract*—This paper introduces a model of affect to improve prosody in text-to-speech synthesis. It operates on the discourse level of text to predict the underlying linguistic factors that contribute towards emotional appraisal, rather than any particular surface emotion itself. The architecture of the model is described and its performance is evaluated on three levels—its predictive accuracy on text, its effect on natural speech and its effect on synthesised speech.**

I. Introduction

From an engineering point of view, spoken language can be divided primarily into a *verbal* component and a *prosodic* component [1]. The verbal component comprises the actual words that are used to communicate. Text-to-speech (TTS) systems use the well-established linguistic methodologies of phonology and phonetics to synthesise intelligible verbal speech. The prosodic component, or *prosody*, is the rhythm, stress, and intonation of speech, and contributes to its naturalness. Prosody is much less understood in linguistics, with most theories advocating a sentence-level prosodic hierarchy that maps morpho-syntactic units to prosodic units of different sizes [2], [3]. The work of [4], [5] and others show that some aspects of prosody are governed by linguistic levels higher than the sentence. In fact, [6] provides acoustic evidence from read and spontaneous speech that confirms this theory, and proposes an expanded prosodic hierarchy that includes higher domains such as discourse.

Discourse is a coherent multi-utterance monologue or dialogue text [7], [8], [5]. It is more than a sequence of utterances, just as an utterance is more than a sequence of words. Explicit and implicit discourse devices signify links among utterances, such as anaphoric relations on the one hand, and discourse topic (or theme) and its progression on the other. *Information structure* is the utterance-internal devices that relate the utterance to its context in the discourse, inter alia its contribution to the topic. More formally, the definition *theme/rheme* distinguishes between the part of the utterance that relates it to the discourse purpose, and the part that advances the discourse. *Background/kontrast* (or *givenness/focus*) distinguishes the parts, specifically words, of the utterance that denote actual content from the alternatives that the discourse context makes available.

Beyond discourse and information structure, another pragmatic influence that regulates prosody is *affect*, or emotion. Affect is probably the most intuitive contributing factor of prosody, yet it is also the most difficult to model. Analysis of positive and negative *sentiments* in text is an easier, yet useful precursor to detecting affect. Research on sentiment analysis and affect detection has explored data-driven and rule-based avenues [9], though it is emphasised that research in affective computing should not be disjunct from emotion theory.

The OCC model [10] is one such theory that takes a step back from the surface level of emotional expressions and rather identifies the underlying factors that contribute towards them. It appraises human emotions from valenced reactions to three aspects of the environment. Firstly, the *consequence* of an event—whether it is desirable or undesirable with respect to one's *goals*. Secondly, the *action* of the agent responsible for the event—whether it is praiseworthy or blameworthy with respect to one's *standards*. Thirdly, the *aspect* of an object—whether it is appealing or unappealing with respect to one's *attitudes*.

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. For example, the event of "I shot the sheriff" may elicit pride over the action if one is an outlaw (one's standard is lawlessness), but fear over the consequence of ending up in jail (one's goal is to remain uncaptured).

[11], [12] implement the OCC model in their TTS system. The accuracy of the natural language processing (NLP) component that predicts the OCC emotions is 80.5% on a 200 sentence test set when the 22 complex OCC emotions are collapsed onto the 6 basic emotions of joy, sadness, fear, anger, disgust and surprise (for comparison to related work). For the speech synthesis component, improvement is shown in the perception of dichotomous sentiment, but the perception of discrete emotions in the synthesised speech still falls far short of those in real speech. This leaves the question of whether a theoretically-motivated approach to modelling affect in synthesised speech is, after all, possible.

Section II will briefly introduce the audiobook as a useful narrative domain for discourse-level analysis in text and speech. In an attempt to improve on the work of [11], [12], a new model of affect will be proposed in Section III that addresses the shortcomings of the initial implementation by operating on the audiobook level. Section IV will relate the experiments on the model and Section V will draw some conclusions about the results.

II. Audiobooks

The text and speech of the audiobook of a novel should be a most suitable source of higher level linguistic and prosodic phenomena. The unfolding plot is directly analogous to a progressively growing discourse context. A knowledge base of the fictional world and its characters is formed by the narrator of the audiobook as he reads out loud. Information in this knowledge base moves from new to given or comes into focus on a continual basis, which should theoretically influence the speech prosody. In the same way the narrator chooses to express affect based on his understanding, or interpretation, of the interaction between the characters and the world and among the characters themselves.

The prototype narrative domain that can be best exploited by a model of affect based on the OCC theory (and for which audiobooks are available) are *children's stories*. These narratives typically have a simpler grammar of English—to boost the accuracy of the NLP—as well as characters of clear distinction between good and evil (protagonists and antagonists)—to boost the accuracy of the OCC model inputs. The Oz series of children's books by L. Frank Baum presents a good case study as it is in the public domain. Electronic versions of the books are obtainable from Project Gutenberg (http://www.gutenberg.org/) (for the text) and LibriVox (http://librivox.org/) (for the audio). The audiobooks to be used as a training set are "The Wonderful Wizard of Oz", "Ozma of Oz", "Dorothy and the Wizard in Oz", "The Road to Oz", "The Patchwork Girl of Oz" and "Rinkitink in Oz". The audiobook for the test set is "The Emerald City of Oz". The typical length of a book is around 40k words/4 hours.

An NLP software package that is most suitable to analyse the audiobook text on a discourse level is Stanford CoreNLP (http://nlp.stanford.edu/software/corenlp.shtml). The accuracies of its most important components are state of the art—POS tagging [13] at 97.24%, constituent parsing [14] at 86.36%, dependency parsing [15] at 80.3% and coreference resolution [16] at 58.3%.

Concerning the audiobook speech on LibriVox, there are two North American English speakers that narrate sizeable subsets of the Oz series. *Phil Chenevert* is a male with an animated, variably toned voice who reads around 21 hours of the training data. *Judy Bieber* is a female with a calmer, evenly toned voice who reads around 12 hours of the training data. Both read around 5 hours of the test data. When a 100 sentence gold standard test subset is singled out (explained later), it comprises around 4 minutes for each speaker.

The phonetic transcriptions of each book are obtained using the Carnegie Mellon University North American English Pronunciation Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). The forced alignment of the audio to the phonetic transcriptions is done with the Hidden Markov Model Toolkit (HTK) [17]. The TTS system Speect [18] processes the NLP and phonetic information to produce synthesised speech with a plugin of the HMM-Based Speech Synthesis System (HTS) engine [19].

III. E-*motif*

A new model, named e-*motif*, will now be put forth in an attempt to improve upon the OCC model implementation of [11]. Its name is a three-fold word play on the important components of this research: (*e*)lectronic *motif*, that is theme, contributes to *emotive* modelling. In other words, e-*motif* takes advantage of the discourse and information structure ("theme") in ("electronic") audiobook text to model affect ("emotion") according to the OCC theory in a more flexible way. This it does by specifying the three cognitive features of *judgment*, *focus* and *tense*, and the three social features of *power*, *interaction* and *rhetoric*.

A. Judgment

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 difficulty lies with the cognitive antecedents. 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. Like [11], e-*motif* aggregates the OCC goals, standards and attitudes into a single sense, or *judgment*, of right and wrong, good and bad. However, it departs from their implementation in that the belief system of the person is purely subjective.

Informally, e-*motif* 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 e-*motif* defines an emotion *anonymously* based on the *interaction* among the underlying semantic variables **A**, v and **P**, and does not commit to their *composition* according to a particular objective belief system. The number of possible affective states produced by e-*motif* is $2^3 = 8$, as illustrated in Table I. The discourse context for the examples in the table is "Policemen are good. Criminals are bad. To save someone is good. To kill someone is bad".

TABLE I
Possible combinations of valenced semantic states

A v **P**	**Gloss**	**Example**
0 0 0	bad A doing bad to bad P	*criminal kills criminal*
0 0 1	bad A doing bad to good P	*criminal kills policeman*
0 1 0	bad A doing good to bad P	*criminal saves criminal*
0 1 1	bad A doing good to good P	*criminal saves policeman*
1 0 0	good A doing bad to bad P	*policeman kills criminal*
1 0 1	good A doing bad to good P	*policeman kills policeman*
1 1 0	good A doing good to bad P	*policeman saves criminal*
1 1 1	good A doing good to good P	*policeman saves policeman*

The implementation of e-*motif* 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. Firstly, right and wrong, good and bad are represented by the boolean values of true (1) and false (0).

The discourse is divided into clauses delimited by verbs—the semantic predicate—that may have a semantic AGENT and/or PATIENT. The AGENT is typically the nominal subject and the PATIENT the direct object, complement or copula.

The good or bad valence of a discourse *entity* (a coreference-resolved semantic AGENT or PATIENT) represented by a noun phrase defaults to the entry of (the lemma of) the head noun in the SentiWordNet lexicon. SentiWordNet [20] assigns a positive or negative sentiment score to each WordNet [21] entry. If no entry is available, a good valence is assigned. e-*motif* follows the methodology of [11] to determine the polarity of a word from SentiWordNet—namely using the net positive sentiment count of all the senses of the particular word found in WordNet—since no WordNet word-sense disambiguation functionality is available in Stanford CoreNLP.

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`). As in [11], 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* (a semantic predicate) 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.

B. Focus

It is very useful to note that, in the linguistic domain, information structure can readily be applied to determine whether the focus of attention in the OCC model lies on either one of the agents and/or objects, or on the event itself. e-*motif* specifies the three focus areas of the consequence for the semantic AGENT, the action of the semantic AGENT (the semantic verb predicate) and the consequence for the semantic PATIENT. These areas can be distinguished *indirectly* based on the interaction among the information status of the discourse entity represented by the AGENT (**A**), the discourse action represented by the verb (ν) and the discourse entity represented by the PATIENT (**P**). The information status is simplified to a *given/new* dichotomy, where a discourse entity or action is given (0) in the current discourse if it is present in the immediately preceding discourse, and new (1) if it is not. Table II shows how the information status values can combine to form proper theme and/or rheme phrase sequences according to [7]. *Importantly, this cognitive feature of focus in e-motif subsumes the prosodic effects of information structure under those of affect.*

TABLE II
TRUTH TABLE FOR THE FOCUS AREAS IN E-*motif*

A ν P	**Information Structure**
0 0 0	$[given\ \ given\ \ given]_{theme}$
0 0 1	$[given\ \ given]_{theme}\ \ [new]_{rheme}$
0 1 0	$[given]_{theme}\ \ [new]_{rheme}\ \ [given]_{theme}$
0 1 1	$[given]_{theme}\ \ [new\ \ new]_{rheme}$
1 0 0	$[new]_{rheme}\ \ [given\ \ given]_{theme}$
1 0 1	$[new]_{rheme}\ \ [given]_{theme}\ \ [new]_{rheme}$
1 1 0	$[new\ \ new]_{rheme}\ \ [given]_{theme}$
1 1 1	$[new\ \ new\ \ new]_{rheme}$

The "current discourse" is defined as the current **AνP** clause and the "immediately preceding discourse" as the previous **AνP** clause. If the coreference-resolved discourse entity in the current AGENT role is found in either one of the previous AGENT, verb or PATIENT roles (a verb can also be an AGENT), then it is marked as given, otherwise as new. The same applies to the discourse action in the current verb role and the discourse entity in the current PATIENT role.

C. Tense

As in [11], e-*motif* models the temporal aspects of the emotions by noting the tense of the verbs in the clauses. The past tense loosely indicates retrospective consequences of the event, present tense the action of the agent and future tense prospective consequences of the event. Negation for disconfirmation of prospects is covered in the valence calculation of the judgment feature. Tense is captured in the POS tags of the verbs as output by Stanford CoreNLP.

D. Power

The social factor of *power* can influence the emotional responses of two interlocutors in a conversation. This is the power, or status, that one interlocutor can have over the other to trigger social dynamics such as authority and submission, for example in parent-child, teacher-student or policeman-criminal relationships.

Now, the narrative of a novel alternates between the indirect speech of the narrator and the direct speech of the characters in the story. In order to capture and make use of this flow computationally, the discourse is grouped into *speech reports*, or turns, each anchored by the direct speech of *one* of the characters. Paragraph structure gives clues to cluster successive statements by the same character, since intermittent indirect speech narratives (usually short) may be present. These narratives, as well as any introductory ones (usually longer), are included in a speech report.

e-*motif* identifies the coreference-resolved SPEAKER (S) and LISTENER (L) of each speech report and determines their good (1) or bad (1) valence through the judgment feature. It then sets up the power feature as illustrated in Table III

TABLE III
POSSIBLE COMBINATIONS OF SPEAKER-LISTENER POWER

S L	Gloss	Example
0 0	bad S speaking to bad L	*criminal said to criminal*
0 1	bad S speaking to good L	*criminal shouted at policeman*
1 0	good S speaking to bad L	*policeman reprimanded criminal*
1 1	good S speaking to good L	*policeman answered policeman*

An interesting by-product of the power feature is that the subjectivity of the judgment feature is additionally refined in case the narrator wants to appraise the emotions of the interlocutors vicariously on their behalf. Suppose the situation where "a bad AGENT does a bad deed to a good PATIENT". If it occurs in indirect speech narrative, the narrator always appraises from his own belief system. However, if it occurs in direct speech dialogue, the narrator has a choice. Suppose he chooses the vicarious option. Then, if a bad SPEAKER is talking about it (admiration/camaraderie), he should sound different to when a good SPEAKER is talking about it (reproach/disassociation). The situation can similarly be extended to differently valenced LISTENERs. All of this can now be modelled.

All text within quotation marks are assumed to be direct speech that forms part of a dialogue. This means that a conversation is always interpreted as between a single SPEAKER and a single LISTENER, with dialogue turns between the two until a new SPEAKER and/or LISTENER is explicitly introduced. The SPEAKER and LISTENER are identified using the following heuristics.

The first sentence in the indirect speech narrative immediately *succeeding* the direct speech in a speech report is searched for a reporting verb. A reporting verb here is a verb that is typically used to introduce direct speech, for example `said`, `shout`, `ask` and `answer`. If that sentence does not contain a reporting verb, then the final sentence in the indirect speech narrative immediately *preceding* the direct speech in the speech report is searched.

The SPEAKER is set to the discourse entity that is the subject of the reporting verb and the LISTENER to the discourse entity that is the indirect object or object of the prepositions `to`, `at` and `of` in a dependency relationship with the reporting verb. If no SPEAKER is found for the current speech report, look in the dialogue turn history and assign the previous LISTENER.

If no LISTENER is found for the current speech report, look in the indirect speech narrative for a discouse entity with whom the SPEAKER interacts. Here, interaction is defined as the LISTENER being the direct object, indirect object or prepositional object of a verb of which the SPEAKER is the subject. If still no LISTENER is found, look in the dialogue turn history and assign the previous SPEAKER, else assume the SPEAKER is talking to himself.

E. Interaction

This feature models the social responses of the characters in their direct speech *interaction* with one another and the environment. *Adaptation* captures the adjustment of a character in response to the environment—it is set for the initial direct speech clause of a character in response to events that occurred in the "environment" of the indirect speech narrative.

Coordination captures the reaction of one character in response to the emotional expressions of another—it is set at each dialogue turn, in other words, for the initial direct speech clause of one character that follows immediately after the final direct speech clause of another character, with no interrupting indirect speech narrative.

Regulation captures the reaction of a character based on his understanding of his own emotional state and relationship with the environment—it is set for each non-initial clause in the direct speech monologue sequence of a character in his dialogue turn.

F. Rhetoric

The name of the feature alludes to "rhetorical question". It is a simple binary feature that distinguishes between statements and questions as a form of *rhetoric*. The main reason for its inclusion is its pronounced effect on sentence-final prosody, namely an F0 downstep for statements versus an upstep for questions.

IV. EXPERIMENTS

The following experimental investigation evaluates the accuracy of e-*motif* in predicting the linguistic features from text and accounting for the prosody in natural and synthesised speech.

A. Affect Detection from Text

In order to test the accuracy of e-*motif*, 100 sentences are selected from the "The Emerald City of Oz" test set. The sentences are strict single AGENT-verb-PATIENT clauses spread over the test set, in order to optimise the semantic preconditions of the model. Each sentence is manually annotated with the correct feature values, where "correct" is not restricted by the correctness of preceding components in the NLP pipeline. In particular, character valences are not determined by copular induction, but assigned on a human intuitive basis according to the protagonistic or antagonistic role of the character in the story. Furthermore, human intuitive coreference resolution is done to track characters in the preceding discourse up to the point of the particular sentence when focus is assigned.

The automatically predicted feature values are compared against the gold standard to produce the accuracies in Table IV. The six features are indicated in normal roman script. The bold "All" signifies all features strictly correct, "Cognitive" signifies all cognitive features (tense, judgment and focus) strictly correct, and "Social" signifies all social features (power, interaction and rhetoric) strictly correct. The italicised "agent, verb, patient, speaker, listener" signify individual role slots within the compound features.

TABLE IV
E-*motif* ACCURACY

Feature	Accuracy (%)			
All				*11*
Cognitive			**15**	
Judgment		41		
- *agent*	*56*			
- *verb*	*81*			
- *patient*	*73*			
Focus		31		
- *agent*	*48*			
- *verb*	*86*			
- *patient*	*71*			
Tense		83		
Social			**47**	
Power		51		
- *speaker*	*63*			
- *listener*	*64*			
Interaction		89		
Rhetoric		100		

The rhetoric feature has a 100% accuracy since it is a direct mapping from the text. Interaction is also a direct mapping, but does not obtain a full score, since the gold standard considered successive direct speech segments in some contexts still to be coordinated, not yet regulated. Tense has a high accuracy due to the well-performing underlying POS tagging algorithm in Stanford CoreNLP. The features of judgment, focus and power, however, all have much lower accuracies because they have compound values that are furthermore dependent on the coreference resolution performance, which is only 58.3% (Section II). In fact, the individual agent, speaker and listener slot accuracies reflect this region. The verb slots score much higher because the verb predicates need not be coreference-resolved—their lemmas are simply considered as canon. The patient slot is in between the agent and the verb slots because the semantic PATIENT can often be an adjectival complement—canonised by lemma—instead of a noun object—which needs to be coreference-resolved. The model performs very poorly when strict correctness of the feature subsets are required, both for the cognitive subset and the social subset, and thus overall.

The next section investigates the acoustic effects of the e-*motif* features in audiobook speech and discuss the ramifications of the low predictive ability of the model.

B. *Affective Prosody in Natural Speech*

[12] examine the changes in speech rate, pitch average, pitch range and intensity when they compare emotional natural speech to neutral natural speech. They evaluate the effects of the six basic emotions of joy, sadness, fear, anger, disgust and surprise. In the case of e-*motif*, the question is asked not about the consequential discrete emotions, but about the antecedental linguistic features.

The modelling adequacy of e-*motif* is evaluated on the aligned natural speech in the audiobook test set, for each speaker, by comparing the means of the distributions of the duration, average F0 and average intensity measures, with and without the linguistic features. The discrete linguistic features need to be binarised in a "one versus many" fashion, *resulting in 30 binary features to be considered.* For each acoustic measure, a *t*-test delivers a verdict on the statistical significance of the difference between the means. The independent two-sample *t*-test statistic for unequal sample sizes and unequal variances is calculated as follows [22]:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1}$$

where μ_1, σ_1^2 and n_1 are the standard sample mean and variance and number of samples in the test set for the distribution with the binary linguistic feature *deactivated.* Correspondingly, μ_2, σ_2^2 and n_2 are for the distribution with the binary feature *activated.*

To test for significance, t is compared to the appropriate t-test table value. The traditional significance level of $p < 0.05$ is adjusted for non-direction (two-tailedness) and Bonferroni-corrected for the 30 binary features to $p < \frac{0.05}{2\times30} \approx 0.001$. The degrees of freedom typically approximate infinity, so the threshold *t*-value is 3.090.

In addition to the *t*-test, a sanity check compares the difference between the distribution means to the *just noticeable difference (JND)*, a threshold for perceptual discrimination. With regard to complex signals such as speech, the JND for duration (tempo) is 5%, for F0 it is 1Hz and for intensity it is 1dB [23], [24].

The acoustics are measured on the phonetic level and only segments that fall under AGENT-verb-PATIENT semantics are considered. The duration values of the segments are available from the alignment information; the F0 and intensity values are extracted with Praat [25].

Most of the activity takes place in the F0 domain. Table V and Table VI list the sample distribution means of the average F0 for each automatically calculated binary linguistic feature when the latter is deactivated ("off") and activated ("on"). The difference ("diff") between the means and its *t*-statistic follow. If the difference is both statistically significant and larger than or equal to the JND, it is highlighted in bold. If it is only significant, it is italicised. If neither, it is normally styled.

In the Phil Chenevert speech (Table V), regarding the judgment features, only judgment$_{011}$ and judgment$_{110}$ have effects that are both statistically significant and perceptually distinguishable, albeit the F0 differences are not much larger than the JND. If a judicial viewpoint by the speaker can be assumed, the two effects might indicate strong cognitive disbelief that motivates extraordinary prosody over the unexpected situations of a bad agent doing a good deed to a good patient and a good agent doing a good deed to a bad patient, respectively. The same surface emotion is not manifested, however, since judgment$_{011}$ results in a lower tone and judgment$_{110}$ in a higher tone. The features of focus$_{011}$, focus$_{101}$ and focus$_{111}$ are prominent (also only just), though for no apparent reasons other than their intended function, except that focus$_{111}$ also indicates a discourse-new clause, and seemingly by a lowering in tone.

All the tense, power, interaction and rhetoric features are statistically and perceptually significant. The contrast between

TABLE V
t-TESTS ON THE MEANS OF THE AVERAGE F0 MEASURE FOR THE AUTOMATIC LINGUISTIC FEATURES, FROM THE *Phil Chenevert* SPEECH OF THE FULL TEST SET (128481 **A**ν**P** SEGMENTS)

Linguistic Feature	**F0 Means (Hz)**			
	off	**on**	**diff**	*t*
judgment$_{000}$	126.428	127.764	1.336	1.120
judgment$_{001}$	126.584	124.640	-1.944	2.515
judgment$_{010}$	126.476	126.334	-0.142	0.198
judgment$_{011}$	127.081	123.121	**-3.960**	**7.369**
judgment$_{100}$	126.415	128.787	2.372	1.745
judgment$_{101}$	126.331	127.504	1.173	1.937
judgment$_{110}$	126.218	128.334	**2.116**	**3.558**
judgment$_{111}$	126.080	126.979	0.899	2.289
focus$_{000}$	126.477	122.166	-4.311	1.087
focus$_{001}$	126.502	125.697	-0.805	0.866
focus$_{010}$	126.305	128.666	2.361	2.956
focus$_{011}$	126.178	127.865	**1.687**	**3.280**
focus$_{100}$	126.437	131.169	4.732	1.722
focus$_{101}$	126.288	129.062	**2.774**	**3.504**
focus$_{110}$	126.423	126.888	0.465	0.686
focus$_{111}$	127.746	125.430	**-2.316**	**5.937**
tense$_{past}$	131.802	122.486	**-9.316**	**23.710**
tense$_{present}$	123.462	131.183	**7.721**	**19.275**
tense$_{future}$	126.003	138.117	**12.114**	**12.432**
power$_{00}$	126.262	130.142	**3.879**	**4.388**
power$_{01}$	125.621	136.993	**11.372**	**14.713**
power$_{10}$	125.510	137.302	**11.792**	**16.777**
power$_{11}$	124.122	136.193	**12.071**	**23.960**
power$_{narrative}$	135.777	120.218	**-15.559**	**39.169**
interaction$_{adaptation}$	125.687	144.235	**18.549**	**17.710**
interaction$_{coordination}$	125.893	138.346	**12.454**	**13.285**
interaction$_{regulation}$	122.900	134.269	**11.368**	**26.958**
interaction$_{narrative}$	135.777	120.218	**-15.559**	**39.169**
rhetoric$_{statement}$	141.659	126.011	**-15.648**	**13.561**
rhetoric$_{question}$	126.011	141.659	**15.648**	**13.561**

TABLE VI
t-TESTS ON THE MEANS OF THE AVERAGE F0 MEASURE FOR THE AUTOMATIC LINGUISTIC FEATURES, FROM THE *Judy Bieber* SPEECH OF THE FULL TEST SET (132870 **A**ν**P** SEGMENTS)

Linguistic Feature	**F0 Means (Hz)**			
	off	**on**	**diff**	*t*
judgment$_{000}$	236.175	244.876	**8.701**	**5.182**
judgment$_{001}$	236.247	238.888	2.641	2.368
judgment$_{010}$	236.119	239.938	**3.819**	**3.737**
judgment$_{011}$	237.085	232.710	**-4.375**	**5.924**
judgment$_{100}$	236.211	245.506	**9.295**	**4.935**
judgment$_{101}$	236.180	238.119	1.939	2.338
judgment$_{110}$	235.979	239.551	**3.571**	**4.372**
judgment$_{111}$	237.830	234.518	**-3.312**	**6.210**
focus$_{000}$	236.467	213.834	**-22.633**	**4.445**
focus$_{001}$	237.056	222.584	**-14.472**	**11.448**
focus$_{010}$	236.523	234.711	-1.812	1.749
focus$_{011}$	236.056	238.096	2.039	2.924
focus$_{100}$	236.457	227.606	-8.851	2.559
focus$_{101}$	236.893	229.083	**-7.810**	**7.418**
focus$_{110}$	236.186	238.609	2.422	2.610
focus$_{111}$	234.525	237.892	**3.368**	**6.338**
tense$_{past}$	237.690	235.414	**-2.275**	**4.264**
tense$_{present}$	236.218	236.684	0.466	0.861
tense$_{future}$	235.937	248.252	**12.314**	**9.419**
power$_{00}$	236.112	241.594	**5.482**	**4.624**
power$_{01}$	235.759	244.220	**8.461**	**8.373**
power$_{10}$	235.593	245.441	**9.848**	**10.074**
power$_{11}$	236.165	237.347	1.182	1.784
power$_{narrative}$	240.777	233.338	**-7.439**	**13.805**
interaction$_{adaptation}$	235.483	255.809	**20.326**	**14.954**
interaction$_{coordination}$	235.536	253.391	**17.855**	**13.708**
interaction$_{regulation}$	236.256	236.717	0.462	0.818
interaction$_{narrative}$	240.777	233.338	**-7.439**	**13.805**
rhetoric$_{statement}$	243.336	236.163	**-7.173**	**4.961**
rhetoric$_{question}$	236.163	243.336	**7.173**	**4.961**

the vocalisation of indirect and direct speech is clear in the effects of the interaction features. The speaker uses a lower tone for indirect speech narrative (represented by interaction$_{narrative}$) than for direct speech dialogue of the story characters (represented by the other interaction features). The rhetoric features correctly model statements with a down-step and questions with an upstep.

Confounding factors are probably present in the tense and power features. The past tense is mostly used in indirect speech narrative (a common writing technique), explaining the decreasing effect on F0 of tense$_{past}$, as opposed to the increasing effect of tense$_{present}$ and tense$_{future}$ in direct speech dialogue. The power features exhibit the same behaviour, since they all occur in direct speech dialogue, except for power$_{narrative}$.

In the Judy Bieber speech (Table VI), the cognitive features show greater cohesion, especially judgment; focus less so, but still more than in the Phil Chenevert speech. Tense, power and interaction behave more or less the same as in the Phil Chenevert case. The features of tense$_{past}$, power$_{narrative}$ and interaction$_{narrative}$ decrease F0 as a result of indirect speech narrative, whereas interaction$_{adaptation}$ and interaction$_{coordination}$ increase F0 to denote direct speech dialogue turns. The features of power$_{00}$, power$_{01}$ and power$_{10}$ behave accordingly. Once again, the rhetoric features perform as expected.

Since the automatically calculated features have a low accuracy (Section IV-A) that can influence the interpretation of the effects, the gold standard features of the 100 sentence test subset are also evaluated. However, they generally confirm the automatic case and do not show any other significant trends.

Despite their poor accuracy, the cognitive features of judgment and focus appear to have a cohesive effect on the natural speech of Judy Bieber, but are only able to model extreme affective states in the Phil Chenevert speech. This is most likely due to the predisposition of Phil Chenevert being more animated in his speech, as compared to Judy Bieber who is calmer. Phil Chenevert displays a type of speaker choice that overpowers the finer prosodic nuances being modelled by *e-motif*.

The social features seem to be more robust, as they fare well across the board. However, whereas the interaction features are explicitly defined to model the differences between indirect speech narrative and direct speech dialogue, these speech phenomena have a confounding effect on the tense and power features.

The next section explores whether the *e-motif* features can be used successfully in speech synthesis, despite their spurious relationships with natural speech.

C. *Affective Prosody in Synthesised Speech*

[12] takes a hand-crafted rule-based approach to model prosody explicitly in their TTS system. The consequential discrete emotions of their model are mapped to acoustic parameters that alter the prosodic behaviour of the system appropriately. Although improvement is shown in the perception

of dichotomous sentiment, the perception of discrete emotions in their synthesised speech do not nearly match those in natural speech accurately enough.

e-*motif* attempts a different route via the HTS framework. The antecedental linguistic features are included in the HTS context labels and the corresponding decision tree questions are defined, in order to model the prosodic effects of e-*motif* implicitly through the data separation process. Table VII lists the format of the HTS labels. The traditional positional and counting features, as suggested by the HTS documentation, are included on the syllable (P context), word (A context), phrase (B context) and clause (C context) levels. They are a naive, but effective way of capturing physiological factors in speech planning—the longer the phrase is in its syllable count, the greater the effort (breath/pitch/energy) is required to realise it; the position of each syllable within the phrase determines what portion of the effort that syllable will receive; et cetera. The labels furthermore contain lexical and phrase stress information. Finally, the e-*motif* cognitive and social features are specified in their own contexts (D and E, respectively).

TABLE VII
FEATURES USED IN THE HTS CONTEXT LABELS

P context: syllable-level phonetic features
p_1: left triphone context (previous phone) p_2: center triphone context (current phone) p_3: right triphone context (next phone) p_4: phone position in syllable: initial, medial, final p_5: phone count in syllable: isolated, short, medium, long
A context: word-level lexical features
a_1: syllable position in word: initial, medial, final a_2: syllable count in word: isolated, short, medium, long a_3: syllable lexical function in word: primary, secondary, none
B context: phrase-level syntactic features
b_1: syllable position in phrase: initial, medial, final b_2: syllable count in phrase: isolated, short, medium, long b_3: word position in phrase: initial, medial, final b_4: word count in phrase: isolated, short, medium, long b_5: word syntactic function in phrase: head, modifier
C context: clause-level semantic features
c_1: syllable position in clause: initial, medial, final c_2: syllable count in clause: isolated, short, medium, long c_3: phrase position in clause: initial, medial, final c_4: phrase count in clause: isolated, short, medium, long c_5: phrase semantic function in clause: agent, verb, patient, other
D context: discourse-level cognitive/pragmatic features
d_1: cognitive/individual clause tense: past, present, future d_2: cognitive/individual clause judgment: 000, 001, 010, 011, 100, 101, 110, 111 d_3: cognitive/individual clause focus: 000, 001, 010, 011, 100, 101, 110, 111
E context: discourse-level social/pragmatic features
e_1: social clause power: 00, 01, 10, 11, narrative e_2: social clause interaction: adaptation, coordination, regulation, narrative e_3: social clause rhetoric: statement, question

Three distinct synthetic voices are trained on the audiobook training set, for each speaker, with the e-*motif* features automatically calculated. A "Baseline" version uses only the P, A, B and C contexts in the HTS labels. A "Cognitive" version adds the D context to the "Baseline" defaults. A "Social" version adds the final E context to the "Cognitive" ones. The contribution of the cognitive and social contexts are separately evaluated because of their unique effects (or non-effects) on natural speech noted in the previous section.

The synthetic voices are successively compared to each other—that is "Cognitive" to "Baseline", and "Social" to "Cognitive", for each speaker—by determining which voice synthesises speech from the text in the full audiobook test set that is *closer* to the original natural speech in the same. Once again, this happens on the phonetic level and only segments that fall under AGENT-verb-PATIENT semantics are considered. The distances between the synthesised and natural segments are calculated for the acoustic measures of duration, F0 and intensity, where the distances for the latter two time-series are represented by their dynamic time warping (DTW) costs (Euclidean distance-based).

The statistical significance of the voice comparisons are determined with McNemar's test, a chi-square test for paired sample data [22]:

$$\chi^2 = \frac{(|n_1 - n_2| - 0.5)^2}{n_1 + n_2} \tag{2}$$

where n_1 is the number of samples in the test set accredited to the first synthetic voice and n_2 to the second synthetic voice. χ^2 has a chi-squared distribution with one degree of freedom (if $n_1 + n_2$ is large enough, which is true for the full test set). To test for significance, χ^2 is compared to the appropriate chi-square table value. For a significance level of $p < 0.05$ and one degree of freedom, the table gives a threshold value of 3.841. If $\chi^2 \geq 3.841$ the synthetic voice with the most votes is significantly closer to the natural voice than the other synthetic voice. If $\chi^2 < 3.841$ the result is insignificant and the two synthetic voices can be said to be similar in closeness to the natural voice.

The results of the synthetic voice comparisons are listed in Table VIII and Table IX. Each table lists the test set sample allocations to the different voices (or "Equal") for the acoustic measures "Duration", "F0" and "Intensity". The last column in the table indicates the χ^2-value for each comparison. If the "Cognitive" voice is significantly closer than the "Baseline" voice or the "Social" voice is significantly closer than the "Cognitive" voice, the entry is highlighted in bold.

TABLE VIII
MCNEMAR COMPARISONS BETWEEN THE SYNTHETIC VOICES ON THE FULL TEST SET, FOR THE *Phil Chenevert* SPEECH

Measure	**A*v*P Segments**				χ^2
	Total	**Baseline**	**Cognitive**	**Equal**	
Duration	128481	49632	49510	29339	0.149
F0	128481	57450	57164	13867	0.711
Intensity	128481	63921	64559	1	3.163
Measure	**A*v*P Segments**				χ^2
	Total	**Cognitive**	**Social**	**Equal**	
Duration	128481	48291	47421	32769	7.899
F0	128481	55841	**59200**	13440	**98.048**
Intensity	128481	63629	**64851**	1	**11.613**

The synthesised voices trained on the Phil Chenevert speech perform as expected. The cognitive features do not contribute significantly enough to the quality of the HTS data separation process, since the e-*motif* judgment and focus features generally have no discernable effect on the Phil Chenevert natural speech, and the tense feature effects are confounded by

TABLE IX

McNEMAR COMPARISONS BETWEEN THE SYNTHETIC VOICES ON THE FULL TEST SET, FOR THE *Judy Bieber* SPEECH

Measure	**AνP Segments**				χ^2
	Total	**Baseline**	**Cognitive**	**Equal**	
Duration	132870	46333	45598	40939	5.868
F0	132870	59704	60173	12993	1.831
Intensity	132870	67532	65299	39	37.522

Measure	**AνP Segments**				χ^2
	Total	**Cognitive**	**Social**	**Equal**	
Duration	132870	45078	44741	43051	1.261
F0	132870	60046	59649	13175	1.313
Intensity	132870	66296	66535	39	0.428

direct speech dialogue factors. The more robust social features, which model the strong differences between indirect speech narrative and direct speech dialogue, do improve the quality.

The Judy Bieber case is different for the worse, since the cognitive version of the synthetic voices is not an improvement over the baseline version, even though the cognitive features have an effect on the natural speech. Furthermore, the social version shows the same quality, despite the social features also being prominent in the natural speech. The HTS framework is most likely smoothing out the finer prosodic nuances in the more evenly toned speech of Judy Bieber, as a consequence of the positional and counting features in the HTS labels that model the speech more robustly than the e-*motif* features during the decision tree clustering process. The strength of these positional and counting features has been noted in a previous study [26].

V. Conclusion

The experiments reveal a few important antitheses in the ability of e-*motif* to model prosodic behaviour in speech. e-*motif* is able to model the prosodic differences between indirect speech narrative and direct speech dialogue via the indirect effects of its social features. Phil Chenevert makes strong use of such prosody, since the effects are significant in his natural speech and impact the HTS data separation process well enough to produce better quality synthesised speech. On the contrary, Judy Bieber appears to moderate her tone in such a way that the naturally significant social features do not influence the quality of her synthesised speech.

e-*motif* is able to model cognitively-based prosody in the evenly toned natural speech of Judy Bieber, but is at a loss in the variably toned natural speech of Phil Chenevert. However, that same even tone is the downfall in speech synthesis, since the computationally much simpler positional and counting features can account for such prosody with similar quality as the complex e-*motif* features do. Since the positional and counting information might be viewed as a naive kind of syntactic structure, the question arises of whether the cognitive features show an effect in the natural speech because of cognition's sake or because of confounding structural factors.

If the latter is true, the implication is then that prosodic phenomena can and need only be robustly explained by *superficial structure* at the current grain of NLP analysis—that is sentence-internal syntactic-like structure, and sentence-external dialogue structure.

References

[1] P. Taylor, *Text-to-Speech Synthesis*, 1st ed. Cambridge University Press, 2009.

[2] E. Selkirk, *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge: MIT Press, 1984.

[3] M. Nespor and I. Vogel, *Prosodic phonology*. Dordrecht: Foris, 1986.

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

[5] 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.

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

[7] M. Steedman, "Information structure and the syntax-phonology interface," *Linguistic Inquiry*, vol. 34, pp. 649–689, 2000.

[8] I. Kruijff-Korbayová and M. Steedman, "Discourse and information structure," *Journal of Logic, Language and Information*, vol. 12, pp. 249–259, 2003.

[9] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions On Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

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

[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] 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.

[14] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423–430.

[15] M. de Marneffe, B. MacCartney, and C. D. Manning, "Generating typed dependency parses from phrase structure parses," in *LREC 2006*, 2006.

[16] 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.

[17] 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.

[18] J. A. Louw, "Speect: A multilingual text-to-speech system," in *Proceedings of the 19th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2008, pp. 165–168.

[19] 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.

[20] 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.

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

[22] S. Boslaugh and P. A. Watters, *Statistics in a nutshell*, 1st ed. O'Reilly Media, Inc., 2008.

[23] H. Quené, "On the just noticeable difference for tempo in speech," *Journal of Phonetics*, vol. 35, no. 3, pp. 353–362, 2007.

[24] B. Kollmeier, T. Brand, and B. Meyer, *Springer Handbook of Speech Processing*. Springer Berlin Heidelberg, 2008, ch. Perception of Speech and Sound, pp. 61–82.

[25] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," http://www.praat.org/.

[26] G. I. Schlünz, E. Barnard, and G. B. van Huyssteen, "Part-of-speech effects on text-to-speech synthesis," in *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2010.