NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT

Advanced natural language processing for improved prosody in text-to-speech synthesis

G. I. Schlünz
22105034

Thesis submitted in fulfilment of the requirements for the degree Doctor of Philosophy in Information Technology at the Vaal Triangle campus of North-West University.

Supervisor: Prof. E. Barnard

May 2014

Acknowledgements

- I thank Prof. Etienne Barnard for his guidance through the course of my studies. He kept me on track in the initial exploratory phase, gave me the freedom to develop my ideas in the middle, and provided helpful suggestions for the final experiments.
- I am grateful to be a part of the Human Language Technology Research Group at the CSIR Meraka Institute. The relaxed work environment and friendly colleagues were pivotal to the success of my studies.

Abstract

Text-to-speech synthesis enables the speech-impeded user of an augmentative and alternative communication system to partake in any conversation on any topic, because it can produce dynamic content. Current synthetic voices do not sound very natural, however, lacking in the areas of emphasis and emotion. These qualities are furthermore important to convey meaning and intent beyond that which can be achieved by the vocabulary of words only. Put differently, speech synthesis requires a more comprehensive analysis of its text input beyond the word level to infer the meaning and intent that elicit emphasis and emotion. The synthesised speech then needs to imitate the effects that these textual factors have on the acoustics of human speech.

This research addresses these challenges by commencing with a literature study on the state of the art in the fields of natural language processing, text-to-speech synthesis and speech prosody. It is noted that the higher linguistic levels of discourse, information structure and affect are necessary for the text analysis to shape the prosody appropriately for more natural synthesised speech. Discourse and information structure account for meaning, intent and emphasis, and affect formalises the modelling of emotion. The OCC model is shown to be a suitable point of departure for a new model of affect that can leverage the higher linguistic levels.

The audiobook is presented as a text and speech resource for the modelling of discourse, information structure and affect because its narrative structure is prosodically richer than the random constitution of a traditional text-to-speech corpus. A set of audiobooks are selected and phonetically aligned for subsequent investigation.

The new model of discourse, information structure and affect, called e-*motif*, is developed to take advantage of the audiobook text. It is a subjective model that does not specify any particular belief system in order to appraise its emotions, but defines only anonymous affect states. Its cognitive and social features rely heavily on the coreference resolution of the text, but this process is found not to be accurate enough to produce usable features values.

The research concludes with an experimental investigation of the influence of the e-*motif* features on human speech and synthesised speech. The aligned audiobook speech is inspected for prosodic correlates of the cognitive and social features, revealing that some activity occurs in the intonational domain. However, when the aligned audiobook speech is used in the training of a synthetic voice, the e-*motif* effects are overshadowed by those of structural features that come standard in the voice building framework.

Key words: natural language processing, text-to-speech synthesis, prosody, discourse, information structure, affect, OCC model, e-*motif*.

Contents

List of Tables

List of Figures

List of Abbreviations

AAC	Augmentative and Alternative Communication
DNN	Deep Neural Network
DP	Dynamic Programming
DSP	Digital Signal Processing
DTW	Dynamic Time Warping
F0	Fundamental Frequency
G2P	Grapheme-to-Phoneme
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
HTS	HMM-Based Speech Synthesis System
IP	Intonation Phrase
JND	Just Noticeable Difference
MFCC	Mel-Frequency Cepstral Coefficients
NLP	Natural Language Processing
NSR	Nuclear Stress Rule
OCC	Ortony, Clore and Collins
POS	Part-of-Speech
PP	Phonological Phrase
PW	Prosodic Word
SAAR	Sentence Accent Assignment Rule
ToBI	Tones and Break Indices
TTS	Text-to-Speech
WDP	Word DP Score
XP	Syntactic Phrase

Chapter 1

Introduction

1.1 Background

A basic need of the human condition is to communicate; yet, for some individuals, this is beyond their natural reach. Enter technology. People with speech impediments can receive a voice for the first time in their lives through a speech-enabled augmentative and alternative communication (AAC) system. Such a system emulates and expedites the language production process for the user by optimising (easing) word and sentence formation and vocalising the result with either pre-recorded human speech or synthesised computer speech (Fossett and Mirenda, 2009). The former option is unfortunately not a very pragmatic solution, because it is a very time consuming process to build up an inventory of recordings. Furthermore, in the construction of a message, it limits the user to the static content that is available in the inventory. The advantage of speech synthesis is that it can create dynamic content. The vocabulary is infinite and any combination of words can be spoken, which is a most desirable feature for proper communication.

Current speech synthesis systems are able to pronounce words in an understandable way, albeit sounding a little robotic. However, speech is more than just its verbal word content; it also serves the higher functions of communication. A speaker typically has a reason why he wants to communicate, that is a purpose or intent. Towards this end, he formulates a message using a choice of words that, together, will convey a certain meaning to the listener in the context of the current topic at hand. He does not have to rely on just this choice of words to promote his intent though. Much of the meaning and intent in everyday human speech can be communicated with non-verbal cues such as emphasis and emotion (Taylor, 2009). Emphasis places more importance on the meaning of certain words than on that of others, whereas emotion adds a dimension of how the speaker feels about the message, for example happy or sad. The challenge of synthesised computer speech is, therefore, not only to *sound* more human-like by incorporating devices like emphasis and emotion, but also to *convey* the meaning and intent that they contribute beyond the word level. This study will explore these concepts for the language of English.

1.2 Problem Statement

The aforementioned two-fold challenge can be described in more detail as follows:

1. Most speech synthesis systems take electronic text as input. On face value, the text is nothing more than a sequence of characters. A premise for good speech output is good text analysis that first of all identifies the words and converts them into a sequence of sounds as pronounced in the particular language. The state of the art utilises pronunciation dictionaries and rules successfully to cover these aspects. However, it is much harder to extract meaning and intent from the text. A deeper analysis

of the interaction among the words will be required to interpret the factors that give rise to emphasis and emotion.

2. Once the factors of meaning, intent, emphasis and emotion are inferred from the text, the synthesised speech needs to be adjusted to reflect the appropriate speech acoustics that convey these devices. It is well known that people vary their speech by speaking more quickly or more slowly, more loudly or more softly, and raising or lowering their tone. However, it is difficult to account for the exact relationship between the textual factors and these acoustic indicators in the speech.

1.3 Research Questions

Following from the above statements, these questions may be asked:

1. How accurately can the factors of meaning, intent, emphasis and emotion be predicted from text using current text analysis tools?

2. Can systematic acoustic correlates of these textual antecedents be determined and applied to speech synthesis (whether manually or automatically assessed)?

1.4 Aims

The research questions lead to the following aims:

1. To build a system that can model meaning, intent, emphasis and emotion with existing text processing resources and techniques.

2. To verify empirically the acoustic phenomena caused by these textual factors in human speech and use them to improve the naturalness of synthesised speech.

1.5 Hypotheses

The following hypotheses are made in answer to the research questions:

1. The computational analysis of natural language text has advanced with great strides, seeing the development of many tools and resources for many languages. It is especially true for English, where analysis has progressed beyond the word level to start looking into language phenomena of interconnected words on the sentence and paragraph levels. Comprehensive dictionaries have also been compiled that specify meaning and other useful knowledge about language, and can be used to great effect. This successful trend thus favours the hypothesis that meaning, intent, emphasis and emotion could be predicted accurately enough from text to be usable for speech synthesis.

2. In a similar fashion, studies into the effects of language on speech patterns shows promising results. Researchers have developed theories that can explain seemingly irregular behaviour on the word level by taking sentence and paragraph factors into account. It is, therefore, prudent to hypothesise that considering meaning, intent, emphasis and emotion should allow for a better model of acoustic behaviour that can be used to produce more natural synthesised speech.

1.6 Contributions

The study hopes to make the following contributions to the scientific community:

1. To release a text analysis system for speech synthesis that can predict suitable acoustics from text by tracking meaning, intent, emphasis and emotion.

2. To provide systematic acoustic correlates of meaning, intent, emphasis and emotion, including an aligned corpus with these annotations.

1.7 Research Methodology

The methodology employed to reach the research aims can be summarised briefly in the following steps:

1. The point of departure is a review of the literature on text analysis and speech synthesis, giving special attention to the relationship between the textual factors of language and the acoustics of speech, in order to be familiarised with the concepts and state of the art in these disciplines. The literature study will also formalise the notions of meaning, intent, emphasis and emotion.

2. An important resource for the development of a model of speech and language is a corpus where the audio of the speech is linked to the text. Traditional collection methods in speech synthesis are not suitable, however, since they only try to cover word-level and sentence-level acoustics. An innovative solution needs to be found to capture paragraph-level acoustics. The quality of the text and audio links will be evaluated with an automatic procedure on the whole corpus and confirmed with manual inspection of a subset.

3. The model of meaning, intent, emphasis and emotion needs to predict these factors automatically from text. The most appropriate related work on text analysis from the literature study must be critically evaluated and its predictive ability extended with novel features to produce a more accurate system. The performance of the system will be tested against a manually annotated gold standard subset of the text data of the corpus.

4. The acoustic correlates of meaning, intent, emphasis and emotion in human speech will be investigated using statistical t-tests on the speech data of the corpus, for the cases when these textual factors are present and absent. The effects on synthesised speech will be evaluated using McNemar comparisons between the cases when the factors are included and excluded in the computer voice building process.

1.8 Chapter Overview

This introduction to the research is Chapter 1 of the thesis. Chapter 2 relates the first step in the research methodology, namely the literature study of text analysis and speech synthesis. For the second step, Chapter 3 introduces the audiobook as a text and audio resource of human speech that operates beyond the word and sentence level. Chapter 4 expounds on the third step in the methodology by evaluating the literature that forms the basis for a new model of meaning, intent, emphasis and emotion. The new model is described and its predictive performance is tested on audiobook text. For the fourth step, Chapter 5 reports on the acoustic experimentation on the audiobook speech, including the synthesis evaluation. Finally, Chapter 6 concludes the research by discussing the findings and recommending future work.

Chapter 2

Literature Study

2.1 Introduction

The research effort commences with a literature study on text analysis and speech synthesis, and how these two disciplines intersect. The various concepts involved are discussed and the state of the art is assessed. The chapter then turns its focus to the specific area of prosody, which is the key to understand acoustic behaviour in human speech and, therefore, the key to produce more natural synthesised speech. The influence of meaning, intent, emphasis and emotion will be investigated before a conclusion is reached on the way forward to address the research aims.

2.2 Natural Language Processing

Natural Language Processing (NLP) is a multi-disciplinary field that borrows from computer science, linguistics and cognitive psychology. It combines their theory with computation to process natural (human) language text. In other words, NLP entails the computational representation and analysis—that is understanding and generation—of the text (Liddy, 2001).

2.2.1 Concepts

NLP processes text on different levels of linguistic analysis (Liddy, 2001):

Phonology This is the study of how speech sounds function and are organised in a particular natural language. Conversely, *phonetics* analyses the physical production of speech, independent of language. Some important terminology are the following: A *phoneme* is the smallest theoretically contrastive unit (able to distinguish words) in the sound system of a language. A *phone* is the smallest physically identifiable unit (yet not able to distinguish words) in speech. A phoneme is realised as one or more phones in different phonemic contexts or environments—these phones are termed *allophones*. For example, the aspirated `[pH]` in `pin` and the unaspirated `[p]` in `spin` are allophones of the phoneme `/p/` (Jurafsky and Martin, 2009).

Morphology The smallest meaningful unit in the grammar of a language is called a *morpheme*. This level then performs morphological decomposition of words into *roots* and *affixes* to infer their internal structure. Consider the example word `misjudged`. A root carries the principal part of meaning in the word, namely `judge`. An affix augments the meaning of the principal part. It can be a *prefix* that is prepended to the word, namely `mis-` meaning "wrong", or a *suffix* that is appended to the word, that is `-ed` indicating the past tense (Jurafsky and Martin, 2009).

Lexicology Lexical analysis determines the underlying meaning or sense of individual words, typically by lookup in a dictionary called a *lexicon* (Jurafsky and Martin, 2009), such as WordNet (Fellbaum, 1999). If a word has multiple senses, it is disambiguated at the semantic level.

Syntax This level infers the grammatical structure of the sentence, that is the structural dependencies among the constituent words. It includes the tagging of the words with Part-of-Speech (POS) categories, for example *noun*, *verb* and *preposition*. The word-POS tag sequences are, in turn, grouped with *constituent* parsing into phrases such as *noun phrases* (headed by a noun), *verb phrases* (headed by a verb) and *prepositional phrases* (headed by a preposition). The structure is most intuitively represented as a tree, of which an example can be seen in Figure 2.1. The grammatical order required of the parts of speech within these structures helps to eliminate the ambiguity of multiple such categories for a single word.

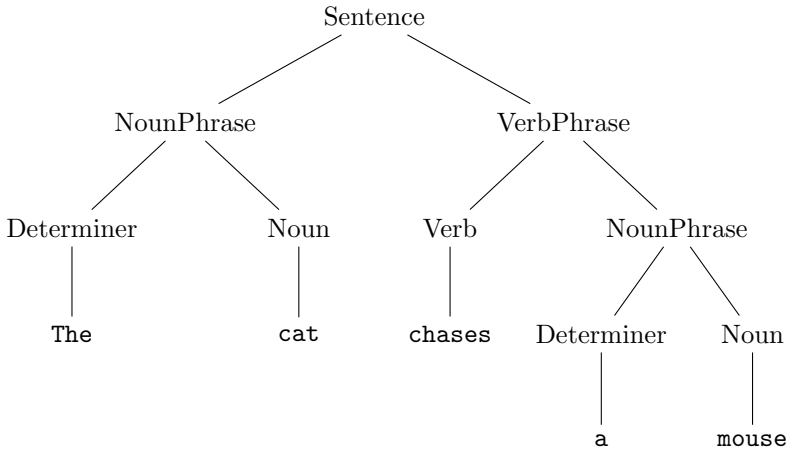


Figure 2.1: The syntactic tree for the sentence `The cat chases a mouse.`

Semantics In general, this is the study of meaning of linguistic expressions. More narrowly defined, it is the study of word sense on the sentence level, not yet considering discourse and pragmatic factors (explanations to follow) (Jurafsky and Martin, 2009). At this level, the meaning of the remaining ambiguous words from the lexical stage are resolved by considering the interactions among the individual word senses in the sentence. This is called *word-sense disambiguation*. The representation of the meaning of words may be done using *first-order logic*. This decomposes a word into its basic properties or semantic primitives. When these primitives are shared among words in the lexicon, meaning can be unified across and inferences drawn from the words (Liddy, 2001). The sentence

$$\texttt{John is the father of Michael.} \tag{2.1}$$

could be represented with first-order logic as follows:

$$\begin{aligned}&\textsc{Relation}(\textsc{Object}(\textsc{Type}(\texttt{father}), \textsc{Agent}(\texttt{John})),\\ &\qquad\textsc{Object}(\textsc{Type}(\texttt{son}), \textsc{Agent}(\texttt{Michael})))\end{aligned} \tag{2.2}$$

The expression Agent is called a *predicate*, which assigns a property to its *argument* `John`. Such *deep* semantics are difficult to determine automatically, so a simpler alternative, *shallow* semantic parsing, is employed in practice. A different view of syntactic parsing, called *dependency* parsing, can be used

to infer shallow semantic properties among words in the sentence in (2.1):

$$\begin{aligned} &\textsc{determiner}(\texttt{father, the})\\ &\textsc{copula}(\texttt{father, is})\\ &\textsc{nominal\_subject}(\texttt{father, John})\\ &\textsc{preposition\_of}(\texttt{father, Michael}) \end{aligned} \tag{2.3}$$

Manually compiled lexicons, such as VerbNet (Kipper et al., 2000) and FrameNet (Baker et al., 1998), also provide semantic structures of common verbs and their arguments.

Discourse Whereas syntax and semantics are, therefore, sentence-level analyses, this level of analysis functions on the whole document or discourse, connecting meaning (for example POS, number agreement, gender, et cetera) across sentences. *Coreference resolution* is a technique that automatically tracks all the *mentions* of a particular discourse entity in a discourse and stores them in an indexed *coreference chain*. The entity is typically represented by the initial common or proper noun phrase and its mentions can be other noun phrases or pronouns. For example, in the sentence:

$$\texttt{Michael is a boy. He likes the girl. She is pretty.} \tag{2.4}$$

two coreference chains are formed: {`Michael`, `a boy`, `He`} on account of the copula relation and male gender agreement, and {`the girl`, `She`} on account of the female gender agreement.

Pragmatics This is the study of meaning in context over and above that which can be captured by the text, for example the intent, plan and/or goal of the speaker, the status of the parties involved and other *world knowledge*. Pragmatics is in this way an explanation of how humans are able to overcome the inherent ambiguity in natural language sentences. Consider the following example:

$$\texttt{The boy hit his little brother. He cried.} \tag{2.5}$$

The coreference resolution of `He` cannot take place on the discourse level because the pronoun agrees in number and gender with both the subject and the object in the first sentence. Pragmatic knowledge that pain inflicted on a young child will normally lead to tears is required to associate `He` with the object `his little brother`.

2.2.2 Software

NLP software have become widely available: some as programs that implement a single process, such as either POS tagging or constituent parsing; others as packages that support a whole pipeline, from POS tagging through constituent and dependency parsing to coreference resolution. The machine learning algorithms that are generally employed in NLP are discussed in detail in Schlünz (2010). The software to be used in this research is Stanford CoreNLP (`http://nlp.stanford.edu/software/corenlp.shtml`), a well-established and supported package that robustly performs all the pipeline functions for English. The accuracies of its most important components are state of the art:

- POS tagging (Toutanova et al., 2003) at 97.24%
- Constituent parsing (Klein and Manning, 2003) at 86.36%
- Dependency parsing (de Marneffe et al., 2006) at 80.3%
- Coreference resolution (Lee et al., 2011) at 58.3%

The cited works document the specific underlying algorithms and their implementations, as well as the data sets against which they were evaluated. In particular, the coreference resolution system was the top ranked system at the CoNLL-2011 shared task—a fact which motivates its choice for this study, as coreference resolution will be shown to be a critical component for the modelling of meaning, intent, emphasis and emotion.

2.3 Text-to-Speech Synthesis

Text-to-Speech (TTS) synthesis is the automatic generation of speech from text. It incorporates NLP to process and annotate the text with useful linguistic information that is needed to synthesise proper speech with digital signal processing (DSP) techniques.

2.3.1 Concepts

Traditionally, TTS comprises the following stages (adapted from Taylor (2009)):

Text segmentation The first stage splits the character stream of the text into initial manageable units, inter alia paragraphs, sentences and tokens. Paragraphs are typically terminated by two or more newline characters. Sentence boundaries are usually marked by punctuation. Tokens, which are the written forms of the unique words yet to be discovered, are often delimited by spaces.

Text decoding The second stage decodes each token into one or more uniquely pronounceable words. Non-standard word tokens such as numbers, dates and abbreviations are classified and expanded into their standard word natural language counterparts in a process called *normalisation*. Examples of expansions are:

`101` → `one hundred and one` (2.6)

`2010/11/19` → `nineteen november twenty ten` (2.7)

`etc.` → `et cetera` (2.8)

A special case of *homograph disambiguation* then disambiguates homographs[1] among the token expansions that are not homophones[2]. Consider the following:

`bear` → BEAR-ANIMAL or BEAR-BURDEN? (2.9)

`bass` → BASS-FISH or BASS-MUSIC? (2.10)

`bear` in (2.9) does not need to be disambiguated, but `bass` in (2.10) does. The classification techniques employed in the normalisation and disambiguation processes range from simple regular expression rules to more elaborate context-sensitive rewrite rules, decision lists, decision trees and Naive Bayes classifiers (Sproat et al., 2001; Taylor, 2009; Yarowsky, 1996).

Text parsing The third stage infers additional lexical, syntactic and morphological structures from the words that are useful for the pronunciation and prosodic modelling stages to follow. The tasks include the assignment of *POS categories* to words, the parsing of *syntactic phrases* and *morphological analysis*, the identification of stems and affixes in words.

Pronunciation modelling The fourth stage models the pronunciation of individual words. It maps the words to their constituent phonemes, either by looking up known words in a *lexicon* or by applying *grapheme-to-phoneme (G2P) rules* to unknown words. *Syllabification* divides the words into syllables.

[1] Words with different meanings but the same written form.
[2] Words with the same pronunciation irrespective of their meaning or written form.

Word-level stress (an inherent property of isolated words: it is stress on certain syllables) or *tone*, depending on the language type, is then assigned to the syllables.

Prosodic modelling The fifth stage predicts the prosody of the whole sentence, namely the *prosodic phrasing* (pauses and intonational phenomena at phrase boundaries), *phrase stress* (a phenomenon of connected speech: certain words in a phrase are stressed according to their word-level stress, at the expense of reducing the word-level stress of the other words) and the melody or *tune* of the entire sentence (for example, questions versus statements).

Speech synthesis The sixth stage encodes the above information into speech using DSP.

In TTS the first five stages of text analysis are collectively referred to as the *NLP frontend*. The sixth stage of speech synthesis is also called the *DSP backend*. The backend can synthesise the speech according to the following major techniques (Taylor, 2009):

Articulatory synthesis This parametric method uses an articulatory model of the human vocal tract[3] to simulate the *physical* process of speech production. The control parameters of the model are (inter alia) sub-glottal pressure, vocal fold tension and the relative positions of the different articulatory organs (Styger and Keller, 1994).

Formant synthesis The vocal tract has certain major resonant frequencies[4] that change as the configuration of the vocal tract changes. The spectral peaks of the resonances are called *formants* and are the distinguishing frequency components of speech (Taylor, 2009). A formant synthesiser thus aims to simulate the *acoustic* process of speech production in a *source-filter* paradigm—the source models the glottal waveform (a pulse train for voiced sounds and random noise for unvoiced sounds) and the filter models the formant resonances of the vocal tract[5] (Smith, 2008a).

Diphone synthesis This is a concatenative synthesis technique that moves away from the explicit rule-based models of speech production towards a data-driven approach for generating speech content. An inventory of segmented units of recorded speech is compiled, one instance for each unique type, and concatenated at runtime to synthesise new speech (Taylor, 2009). Intuitively, phoneme waveforms would make sense as the units, but their concatenation poses problems due to coarticulation effects. *Diphone* waveforms, which capture the transition from the middle of one phoneme to the middle of the next, provide a workaround since there is minimal coarticulation at their boundaries (Styger and Keller, 1994). The original diphone units available to the synthesiser will not have the required prosody of the target utterance, so DSP techniques are used to modify the neutral pitch and timing of these diphones to match those of the specification (Taylor, 2009).

Unit-selection synthesis Unit-selection is like diphone synthesis but uses a much larger inventory of multiple units per unique type, recorded in different prosodic contexts (not only pitch and timing, but also stress, phrasing, et cetera). At runtime, the synthesiser selects the most appropriate sequences of units that fit the specified prosodic criteria, according to a *target cost*. In this way the prosody is modelled implicitly by the data, meaning that the quality of the synthesis is heavily dependent on the quality and coverage of the unit database. The DSP stage mostly only has to join the units. However, the joining is not that trivial anymore, since the variability in units necessarily results in

[3] The vocal tract is the cavity through which a sound wave travels—from the glottis (the space between the two vocal folds in the larynx) through the pharynx (the throat) to the lips and nose (Taylor, 2009).

[4] Resonance is the phenomenon of an acoustic system to vibrate at a larger amplitude than normal when driven by a signal which frequency approximates the natural frequency of vibration of the system. Multiple resonant frequencies may be present at the harmonics (frequencies that are an integer multiple) of the natural frequency (Taylor, 2009).

[5] A filter is a system that alters the signal passing through it. This is exactly what the vocal tract does to the sound wave originating at the glottis (Taylor, 2009).

variability at the unit edges—a consideration that is taken into account in the *concatenation cost* (Black et al., 2007; Taylor, 2009).

Limited-domain synthesis For some applications the range of utterances to be synthesised is limited such that it becomes feasible simply to concatenate whole words or phrases from an inventory. When the vocabulary is out-of-range the synthesiser will then fall back on diphone or unit-selection databases. The task is, therefore, to maximise the quality of the most common utterances and have it degrade gracefully to the less common ones (Taylor, 2009).

Hidden Markov Model-based synthesis This is an example of statistical parametric synthesis that borrows concepts from both parametric formant and data-driven concatenative synthesis. It uses the source-filter paradigm to model the speech acoustics, but this time the parameters are estimated from the recorded speech instead of being hand-crafted. During training of the system, both *excitation* (inter alia fundamental frequency; explanation to follow) and *spectrum* (inter alia mel-frequency cepstral coefficients (MFCCs)[6]) parameters are extracted from the data and modelled by context-dependent Hidden Markov Models (HMMs). The contexts considered are phonetic, linguistic and prosodic. Furthermore, each HMM has state *duration* probability densities to model the temporal structure of speech. During synthesis, the duration, excitation and spectrum parameters are generated from the concatenated HMMs, which, in turn, have been selected by a decision tree based on their context. The latter two sets of parameters are used in the excitation generation and synthesis filter module to synthesise the speech waveform (Black et al., 2007).

2.3.2 Software

Speect (Louw, 2008) is a frontend and backend TTS system that will be used in this research. Currently, the frontend processes text on a per-utterance basis up to the syntactic level, but the modular design of Speect allows it to communicate with the Stanford CoreNLP pipeline to perform the higher-level linguistic processing. The backend synthesises speech using the HMM-Based Speech Synthesis System (HTS) (Zen et al., 2007). The HMM models are trained with a slightly altered version of the demonstration script that is provided with the HTS software package.

2.4 Speech Prosody

From an engineering point of view, spoken language can be divided primarily into a *verbal* component and a *prosodic* component (Taylor, 2009). The verbal component comprises the actual words that are used to communicate. State-of-the-art 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 many theories trying to explain the natural phenomena, as well as predict it from text. Consequently, TTS systems do not yet handle it in a systematic way; currently, most (data-driven) systems simply rely on large amounts of data to model the prosodic effects implicitly. The discussion to follow will concentrate on matters of prosody mostly pertaining to English.

2.4.1 Prosodic Modelling

Prosody can be approached from a low-level physical perspective and a high-level theoretical, linguistic perspective. *Duration*, *fundamental frequency (F0)* and *intensity* have been shown to be acoustic correlates

[6] Simplistically, MFCCs are derived in a process that involves the Mel-cepstrum (the spectrum of a non-linearly mapped spectrum) of a signal—they are a good approximation to the response of the human auditory system (Elsabrouty, 2006).

of prosody (Dong et al., 2005; Waibel, 1988). Duration is simply the temporal length of segments of speech, whether it be phones, syllables, words or phrases. Every periodic signal, such as that which speech approximates, has a fundamental frequency given by the inverse of its period (Smith, 2008b). When considering resonance in speech, the fundamental frequency can be likened to the natural rate of vibration of the vocal cords. Terminologically, *pitch* is the perception of fundamental frequency that includes some errors and non-linearities (Taylor, 2009). Intensity is the power of the speech signal normalised to the human auditory threshold and is related to the amplitude of the vocal cord vibrations. *Loudness* is the perception of intensity and is also non-linear (Jurafsky and Martin, 2009).

Theories of prosody have been shaped by the concept of prosodic hierarchies, as introduced by Selkirk (1984) and Nespor and Vogel (1986). These hierarchies capture the insight that morpho-syntactic units are mapped to prosodic units of different sizes, even though this relationship is not completely straightforward (Féry, 2009; Taylor, 2009). At the bottom of the hierarchy a grammatical word typically forms a *prosodic word* (PW) carrying lexical (word-level) stress. At the top of the hierarchy a whole sentence corresponds to an *intonation phrase* with a clear intonational tune (Liberman and Pierrehumbert, 1984). There is less agreement on the intermediate prosodic domains that should be mapped from syntactic phrases, with most researchers assuming two levels of prosodic phrasing such as *minor phrase* and *major phrase* (Selkirk, 1986). Recent works by Féry (2009, 2010), however, advocate that prosodic structure is thoroughly recursive just like syntax and propose intermediate domains called *p-phrases* (PP, for phonological phrases) that can be embedded into one another. Likewise, the domains of prosodic word and *i-phrase* (IP, for intonation phrase) can be recursive.

Certain aspects of the acoustic realisation of prosodic phrases are already well-known, such as the global downstep in the F0 contour and the temporal lengthening of segments towards the end of a phrase (Taylor, 2009; Féry, 2009). Phrase boundaries are thus typically identified by pauses and/or F0 and duration contrasts because of the downstep of the previous phrase and the reset of the following phrase (Féry, 2009, 2010; Tseng, 2010). The local effects on the F0 contour and segment durations caused by phrase stress are less understood. *Phrase stress* (sentence-level stress, or prominence) is the term used for the stress patterns associated with prosodic phrases. It emerges because of the interaction among prosodic words in a phrase and is, therefore, beyond lexical stress (Truckenbrodt, 2006). Phrase stress manifests itself as *pitch accents*, which are significant highs or lows in the F0 contour that emphasise a segment of speech (Taylor, 2009). Prosodic phrasing thus correlates with pitch accents: Féry (2009) puts it simply that every prosodic phrase is assumed to have a head, which is instantiated as a pitch accent, and every pitch accent is the head of a prosodic phrase.

In her seminal study, Pierrehumbert (1980) developed a model of intonation that includes a grammar describing pitch accent and phrasal tone patterns, as well as an algorithm for calculating such contours from the symbolic notation. The well-known ToBI (Tones and Break Indices) annotation system, which has been adapted to many languages, borrows from this work (see Beckman et al. (2006) for a historical overview). Tilt is another model of intonation, but being intended for engineering purposes, it uses a set of continuous parameters instead of abstract categories to model the contours (Taylor, 1992, 2000). MOMEL/INTSINT is a suite of algorithms that models prosody on a *phonetic* level—with quantitative values directly related to the acoustic signal—and on a surface *phonological* level—as a sequence of discrete symbols; a prosodic equivalent to the International Phonetic Alphabet (Hirst, 2005). The assumption of the model is that a raw F0 curve is composed of an interaction between a global intonation pattern, or *macroprosodic* component, and a sequence of constituent phones, or *microprosodic* component, that causes local deviations from the global intonation pattern. MOMEL factors the F0 curve into these two components automatically. It produces a sequence of target points (<time, frequency> pairs) where the first derivative of the curve is zero (usually a turning point). The target points, when interpolated by a quadratic spline function, define the macroprosodic component sufficiently. INTSINT transcribes the MOMEL output as a sequence of abstract

tonal segments.

2.4.2 Prosodic Prediction

Consider the following example of an organisation of prosodic phrases and phrase stress, where the former are delimited by round brackets and the latter is marked with an "x":

```
(                                                    x ) IP
(                          x ) (                     x ) PP
(          x     ) (       x ) (           x ) (     x ) PP
(     x) (x     ) (       x ) (x ) (  x ) (x ) (     x ) PW
The lead vocalist of the band sang a love song to her fans.
```
(2.11)

Several approaches have seen the light to predict prosodic phrasing and phrase stress from syntactic structure. Early theories of phrase prediction are *relation-based* (Nespor and Vogel, 1986) and *edge-based* (Selkirk, 1986), whereas a more recent one is *alignment* in optimality theory (Truckenbrodt, 1999; Selkirk, 2000). Truckenbrodt (1999) defines the constraint *Align-XP,R* which states that "the right edge of each syntactic phrase (XP) must be aligned with the right edge of a p-phrase". It augments this with *Wrap-XP* to prevent XPs from being split up into multiple p-phrases: "for each XP there must be a p-phrase that contains the XP".

Predicting phrase stress starts off by noting that phrase stress on a word always falls on the lexically stressed syllable(s) of the word (Liberman and Prince, 1977). In the early days Chomsky and Halle (1968) proposed the simple *Nuclear Stress Rule* (NSR) that assigns stress on the right of a phrase, cyclically up the hierarchy. The NSR has a couple of shortcomings though, including being unable to assign more than one phrase stress instance in a sentence. For this reason, Gussenhoven (1983, 1992) developed the *Sentence Accent Assignment Rule* (SAAR) whereby "(within a focus) every predicate, argument and modifier must be accented (receive phrase stress), with the exception of a predicate that stands next to an accented argument". The SAAR has been generalised by Truckenbrodt (1999) to *Stress-XP*: "each XP is assigned a beat of phrasal stress". Final p-phrase stress is then strengthened on the i-phrase level (Selkirk, 1995).

Kratzer and Selkirk (2007) give a more principled interpretation of the syntax-phonology interface by using *phase-based spellout* in the minimalist framework of Chomsky (2000, 2001). Kratzer and Selkirk (2007)'s application of this framework developed from the need to explain why there is a difference between the conditions for verbs and their arguments to carry phrase stress. According to the framework, *phases* provide a new mechanism for syntactic derivation. A phase abstracts syntactic structure with a head and a complement. Within a phase, lexical material, which can include a syntactic phrase in the traditional sense, is inserted and constituents may move up to higher phase-internal syntactic positions. At the end of a phase, the material in the complement to the phase head is spelled out. During spellout lower-order phonological form is assigned to words, as well as higher-order prosodic structure (prosodic phrasing and phrase stress). In order to account for the seemingly non-deterministic behaviour of the verbs in a deterministic way, Kratzer and Selkirk (2007) successfully show that, if the verbs are moved into higher-level syntactic and semantic positions, they can be spelled out correctly. In other words, some aspects of prosody are governed by linguistic levels higher than the sentence. In fact, Tseng (2010) provides acoustic evidence from read and spontaneous speech that confirms this theory, and proposes an expanded prosodic hierarchy that includes higher domains such as paragraph and discourse.

2.4.3 Information Structure

Information structure is one such example of a higher linguistic level that has been researched extensively. In order to define information structure, it is necessary to review *discourse* in more depth. A coherent multi-utterance monologue or dialogue text is a discourse (Kruijff-Korbayová and Steedman, 2003). Discourse 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 then the utterance-internal devices that relate the utterance to its context in the discourse. It includes the contribution of the utterance to the discourse topic, but also the knowledge, beliefs, intentions and expectations of the discourse participants. More formally, the definition *topic/comment* or *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.

Krifka (2007) views information structure from a slightly different perspective that helps with its understanding. The notion of a *common ground* in communication (Karttunen, 1974; Stalnaker, 1974) is used, that is the information mutually known between the participants that is to be shared and continually updated or modified as the conversation progresses. Topic then identifies the (semantic) denotation of an expression about which the new information in the comment should be stored in the common ground. Givenness indicates that the denotation is already present in the common ground. Focus, as in the first description, indicates the presence of alternatives in the common ground that are relevant for the interpretation of an expression.

Revisit the sentence in (2.11). If it is preceded in the discourse by:

`The band consisted of a lead vocalist, drummer, pianist and guitar player.` (2.12)

then the information structure of the sentence may be organised as follows:

- The topic is `The lead vocalist of the band`.
- The comment is `sang a love song to her fans`.
- Within the topic, `lead vocalist` has focus and
- `band` is given.

Steedman (1991, 2000, 2007) claims that information structure is indicated by intonation: certain phrasal tunes, as defined by Pierrehumbert (1980), characterise themes and rhemes, respectively. Within both the theme and rheme, the presence of pitch accents identifies words that are in focus. Féry and Ishihara (2009) also explore the prosodic realisations of information structure and perform production experiments to substantiate them. Leaving out the dimension of topic and comment, the work concentrates on focus and givenness. The prosodic domain of an utterance is defined in terms of register lines that propagate down the prosodic hierarchy, so in other words, the F0 contour subsections of phrases, words and syllables can rise and fall, or "shift", within certain registers. An utterance completely new to the discourse is assigned default phrase stress that maps to some prosodic domain (as discussed earlier). It is then shown that focus on a word enlarges its F0 register and givenness compresses it, and the height of pitch accents and boundary tones are adjusted accordingly.

The setup of these experiments, as well as those of other studies, for example Féry (2009) and Selkirk (2007), may be considered somewhat artificial—the discourse context of an utterance is fabricated by preceding the utterance with a single question that induces the utterance as answer. True multi-utterance discourse with a proper thematic progression is not taken into account.

2.4.4 Affect

Beyond information structure there are other pragmatic influences regulating the prosody of discourse, such as speaker intent and *affect*, or emotion. Affect is probably the most intuitive contributing factor of prosody, yet it is also the most difficult to model. Only recently did the work of Picard (1997) popularise the field of affective computing. This field covers a broad spectrum of computational techniques to model emotion from various modalities, such as facial expressions, speech and text (for an excellent review, see Calvo and D'Mello (2010)). In the context of TTS synthesis, affect *detection from text* and *parametrisation into speech* are the topics of interest. 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. Simple approaches use manually compiled lexicons, such as SentiWordNet (Esuli and Sebastiani, 2006) and WordNet-Affect (Strapparava and Valitutti, 2004), to do keyword spotting in texts. Word-level analyses fail, however, where the polarity or emotion of a sentence is determined by the interaction of words on a syntactic and semantic level. An example is:

```
I could not kill the terrorist.
```
(2.13)

where the negative sentiments of `kill` and `terrorist` need to be composed to give a positive connotation, which is then reversed by the negator `not` into a negative emotion of disappointment. Machine learning approaches try to cater for such instances by defining not only lexical features, but also polarity reversal and n-gram or syntactic features (Yu and Hatzivassiloglou, 2003; Wilson et al., 2005). Rule-based approaches typically start out with affective lexicons to assign prior values to a subset of words and then use relational lexicons such as WordNet (Fellbaum, 1999) and ConceptNet (Liu and Singh, 2004) to expand the subset by incorporating syntactically and semantically similar words (Shaikh et al., 2008b). Finally, both approaches incorporate compositional semantic models to combine the prior values of constituent words into a sentence-wide score (Choi and Cardie, 2008; Neviarouskaya, 2010).

Calvo and D'Mello (2010) emphasise the fact that research in affective computing should not be disjunct from emotion theory. Shaikh et al. (2009a) demonstrate this by applying the cognitive theory of Ortony et al. (1988), also known as the *OCC model*, to affect detection. Simplistically, the OCC model states that human emotions are valenced (positive or negative) reactions to three aspects of the environment:

1. Events of concern to oneself
2. Agents that one considers responsible for such events
3. Objects of concern

A shortcoming of the cognitive approach is that it does not consider the social factors of emotion. Calvo and D'Mello (2010) mention five important social processes that influence emotion:

1. Adaptation—adjustments in response to the environment
2. Coordination—reactions in response to emotional expressions by others
3. Regulation—reactions based on one's understanding of one's own emotional state and relationship with the environment
4. Cultural guidelines for the experience and expression of emotions
5. Power (or status) of one party over another

The influence of affect on the acoustic correlates of prosody and speech in general has been observed in many studies. For an overview, see Murray and Arnott (1993) and Schröder (2009). Intense emotions such as fear, anger and joy typically result in faster, louder and more enunciated speech with strong high-frequency energy. Moderate emotions such as sadness and boredom are associated with slower, low-pitched speech with little high-frequency energy (Pollermann and Archinard, 2002). Most researchers agree that the acoustic dimensions of general prosody—duration, F0 and intensity (Section 2.4.1)—are also applicable to affect (Schröder, 2009).

Emotional TTS systems have been investigated by Shaikh et al. (2008a, 2009b, 2010). It was noted through analytical and perceptual tests that commercial TTS voices at the time were not able to synthesise emotions effectively. In an effort to remedy the situation, the authors initially applied their sentiment detection work (Shaikh et al., 2008b) and subsequently their OCC model-based affect detection work (Shaikh et al., 2009a) to TTS. Their setup explicitly adjusts the parameters of speech rate, pitch average, pitch range, pitch (slope) change and intensity to reflect the detected sentiments and emotions. Improvement was shown in the perception of dichotomous sentiment, but the perception of discrete emotions in the synthesised speech still fell far short of those in real speech. This leaves the question of whether the acoustic modelling of affect in speech is, after all, tractable.

2.5 Conclusion

This chapter briefly reviewed the fields of NLP and TTS before expounding in more depth the issue of speech prosody and its linguistic antecedents. It was argued that prosody can really only be modelled appropriately if one climbs the higher rungs of the linguistic hierarchy, namely discourse, information structure and affect. State-of-the-art NLP software can track discourse and information structure with shallow semantic (dependency) parsing and coreference resolution, though do not yet employ these higher linguistic levels towards affect detection. Currently, affect is mostly predicted using (subsentential) lexical, syntactic and semantic devices. The next chapter will motivate the use of the audiobook as a source of discourse-level linguistic and prosodic phenomena, and analyse the alignment of its text and speech.

It was also noted that emotion theory should be consulted in the construction of models of affect. One line of research (Shaikh et al., 2009a) has implemented the OCC model (Ortony et al., 1988) in the linguistic domain. Chapter 4 will evaluate the model and its implementation, and identify their strengths and weaknesses. Since the implementation is not effective enough at modelling affect in speech (Shaikh et al., 2010), a new model will be proposed that leverages the discourse and information structure of audiobook text, not only to account for the cognitive factors, but also the social factors mentioned in the previous section. Chapter 5 will then investigate the effects of the new model on the acoustics in audiobook speech, which should be more amenable to prosody on the discourse, information structural and affective levels.

Chapter 3

Aligned Audiobooks as a Resource for Prosodic Modelling

3.1 Introduction

The previous chapter concluded that most state-of-the-art approaches towards prosodic modelling from text do not employ the higher linguistic levels of discourse, information structure and affect. The works of Féry and Ishihara (2009), Féry (2009) and Selkirk (2007) provide empirical results on how information structure shapes prosody, but their experimental setups only simulate discourse context. Proper thematic discourse is not considered. What complicates the matter further in the TTS community, is that the traditional method of training data collection, where sentences are selected from random source texts to achieve phonetic coverage, is not conducive to elicit true discourse-level prosodic phenomena. This chapter will introduce the audiobook as a hypothesised solution that addresses these concerns. It will describe an automatic alignment process to link the text and acoustics, and comment on the quality of the alignment for the prosodic investigation in a subsequent chapter.

3.2 Motivation

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 (a monologue equivalent of common ground) of the fictional world and its characters is formed by the narrator of the audiobook as he files comments under topics while reading out loud. Information in this knowledge base thus moves from new to given or comes into focus on a continual basis, which should theoretically influence the prosody of the narrator's speech. 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 mostly obtainable from Project Gutenberg (`http://www.gutenberg.org/`) (for the text) and LibriVox (`http://librivox.org/`) (for the audio). On LibriVox there are two North American English speakers that narrate sizeable subsets of the series. *Phil Chenevert* is a

male with an animated, variably toned voice who reads the following books chosen as his training set: "oz01: The Wonderful Wizard of Oz", "oz03: Ozma of Oz", "oz04: Dorothy and the Wizard in Oz", "oz05: The Road to Oz" and "oz07: The Patchwork Girl of Oz". *Judy Bieber* is a female with a calmer, evenly toned voice who reads these books as her training set: "oz03: Ozma of Oz", "oz04: Dorothy and the Wizard in Oz" and "oz10: Rinkitink in Oz". Both speakers read the test set book: "oz06: The Emerald City of Oz".

3.3 Alignment

The audiobook alignment process works as follows. The texts of the various audiobooks are divided semi-automatically into chapters to match the accompanying chapter-level audio files. The division is easily done with regular expressions that match the manually inspected chapter heading formats. For each book, the chapter-level text is processed by the Speect frontend. It detects sentence boundaries, using regular expressions that match sentence-final punctuation, and it produces phonetic transcriptions, using a lexicon and G2P rules that are based on the Carnegie Mellon University North American English Pronunciation Dictionary (`http://www.speech.cs.cmu.edu/cgi-bin/cmudict`). The G2P algorithm is Default&Refine, which is described and evaluated in detail in Davel and Barnard (2008).

The Hidden Markov Model Toolkit (HTK) (Young et al., 2006) is used in the forced alignment of the audio to the phonetic transcriptions of each book. Firstly, one book from each speaker is selected in a seed phase—"oz01: The Wonderful Wizard of Oz" for Phil Chenevert and "oz04: Dorothy and the Wizard in Oz" for Judy Bieber. The chapter-level audio of these two books is aligned to the chapter-level transcriptions using North American English triphone acoustic models trained on the English Broadcast News Speech corpus (Graff et al., 1997). The chapter-level alignments are split at the utterance-level, quality controlled and then used to train speaker-specific triphone acoustic models (the utterance chunking has to be done to make the quality control computationally tractable). The second phase sees all of the audiobooks aligned with the speaker-specific seed models and, again, split into utterances and quality controlled. Time did not allow the training of target data models for each of the audiobooks, but the speaker-specific models are assumed to be sufficient for the task at hand, since the seed and target data audiobooks are all in the same domain. The scripts that perform the audiobook alignment are based on the work of van Heerden et al. (2012).

3.3.1 Automatic Evaluation

The quality control employs the phone-based dynamic programming (DP) technique of Davel et al. (2012) to score the phonetic alignments. Basically, it computes the confidence score of an utterance as the lowest DP cost when aligning the freely decoded phone string to the forced alignment of the provided transcription. In particular, Davel et al. (2012) specify 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 a forced alignment if the target phone string is a segment within a larger utterance, produces a *reference string*.

3. A standard DP algorithm with a pre-calculated scoring matrix is used to align the observed and reference string with each other. The scoring matrix species the cost associated with a specific substitution between a phone in the reference string and the observed string.

4. The resulting score obtained from the best DP 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.

Example output of the quality control per utterance is provided in Figure 3.1. It is basically a list of the words in the utterance (column 1), including possible insertions, and their individual DP scores (column 6) and alignments (column 7 for the reference/aligned string and column 8 for the observed/decoded string). The insertions are caused by freely decoded phones that exceed the number of aligned phones. It can be seen from the entries of "one", "kept" and "his" that a perfect DP alignment of a word receives a word DP (WDP) score of zero. Errors are indicated by scores lower than zero.

```
every    data_000002   2200000   6200000  5  -0.700  [eh v - er iy]      [eh v d r iy]
one      data_000002   6200000   8900000  3   0.000  [w ah n]            [w ah n]
<ins>    data_000002   8900000   8900000  1  -0.500  [-]                 [d]
kept     data_000002   8900000  11800000  4   0.000  [k eh p t]          [k eh p t]
away     data_000002  11800000  14000000  3  -1.333  [ah w ey]           [er w ih]
from     data_000002  14000000  16300000  4   0.000  [f r ah m]          [f r ah m]
him      data_000002  16300000  19500000  3  -1.167  [hh ih m]           [- eh m]
even     data_000002  19500000  23500000  4  -0.500  [iy v ih n]         [iy v ah n]
<ins>    data_000002  23500000  23500000  2  -0.500  [- -]               [d ng]
his      data_000002  23500000  26600000  3   0.000  [hh ih z]           [hh ih z]
<ins>    data_000002  26600000  26600000  1  -0.500  [-]                 [jh]
chief    data_000002  26600000  30800000  3  -1.333  [ch iy f]           [sh iy th]
steward  data_000002  30800000  35800000  6  -0.250  [s t uw - er d]     [s t uw w er d]
<ins>    data_000002  35800000  35800000  1  -0.500  [-]                 [t]
kaliko   data_000002  35800000  42200000  7  -1.071  [k ae l ih k - ow]  [k ah l iy k ah l]
```

Figure 3.1: Example quality control output

Table 3.1 shows the overall alignment statistics on the Phil Chenevert audiobooks and Table 3.3 those on the Judy Bieber audiobooks. They list three broad columns per book that contain information on "All" of the utterances before quality control, on the "Good" utterances that passed the quality control and on the "Bad" utterances that failed the quality control, respectively. The particular information is the number of utterances and audio length, along with the average number of words ("W"), the average number of phones ("P") and the average word DP score ("WDP") per utterance. The subtotals for the training set of books ("train") and the total overall ("all"), which includes the test set, are also given.

To understand the criteria for the quality control[1], it is necessary to look at the distribution of average WDP scores over the audiobooks in Figure 3.2 (for Phil Chenevert) and Figure 3.3 (for Judy Bieber). The histograms in subfigures (a) show that the greater concentration of utterances obtain an average WDP score of more than -0.75. Assuming that the narrator reads the text correctly more often than not, it is reasonable to judge the -0.75 score as a safe threshold to discard the outliers. Subfigures (b) illustrate the effect of utterance length on the WDP scores by plotting the scores against the number of phones in the utterances. The peaks are around a WDP score of -0.25 and 20 phones per utterance.

The supposed phone length of an utterance can be problematic for alignment in two extreme cases. On the one hand, very short utterances are typically interjections that are voiced with extraordinary prosody (for example, "Help!"). On the other hand, very long insertions (as indicated by the DP alignment) may indicate words in the speech that are not in the text. Hence, the quality control criteria are set to the following (in order):

1. If an insertion in the utterance has a length greater or equal to 10 phones, discard the utterance.

2. If an utterance has a length less than 10 phones, discard the utterance.

3. If an utterance has an average WDP score less than -0.75, discard the utterance.

[1] van Heerden et al. (2012) employ the DP timing discrepancies to measure alignment accuracy, but here time did not allow for an investigation into this method, so the simpler threshold alternative is used.

4. Else, keep the utterance.

Table 3.1 thus shows that 17249 utterances (26h16m) are kept out of the total number of 21941 (29h34m) after quality control on the Phil Chenevert audiobooks. Table 3.3 indicates that 11620 utterances (16h54m) out of a total of 13883 (18h50m) pass the quality control on the Judy Bieber audiobooks. Table 3.2 and Table 3.4 give a detailed breakdown of the bad alignment statistics for Phil Chenevert and Judy Bieber, respectively. The average WDP scores for the "Too Long Insertion" and "Too Few Segments" utterances all turn out to be above the threshold of -0.75[2]. This necessitates manual verification of the alignments to test the need for these explicit criteria in the quality control process.

3.3.2 Manual Verification

It is prudent to test not only the quality of the alignments with "Too Long Insertion" and "Too Few Segments", but also the alignment quality across the -0.75 WDP threshold to obtain a discriminatory intuition of the WDP scoring scale. Towards this end, a subset of the "oz06: The Emerald City of Oz" test set of alignments is manually inspected for errors. More specifically, for each quality control criterion, 20 random utterances that intersect both speakers are selected and checked for gross word boundary errors. By *gross* is meant word boundaries that are misaligned between speech and text in a manner worse than that which can be caused by inter-word coarticulation effects. In the gross case, the alignment process will often only be able to recover after a few subsequent words.

In the tables to follow, for each utterance, its order in the test set ("No") and its body text ("Text") is given, along with the average WDP score ("WDP") for both Phil Chenevert and Judy Bieber. The number of gross word boundary errors for each speaker is indicated with an "E". Meta information ("Meta") about extraordinary speaking style or content, if present, is noted. These include voice impersonation of story characters ("person"), animated speech "animate", questions ("question") and extra speech content in chapter headers ("ch start") and footers ("ch end"). A chapter header typically contains the chapter heading and a LibriVox disclaimer. A chapter footer simply signals the end of the chapter. Finally, word substitution by the speaker that did not cause word boundary errors per se is indicated with "$n\times$ sub".

Table 3.5 lists 20 random utterances that obtained good alignments for both speakers, in other words with an average WDP score greater than or equal to -0.75. Only a single word boundary error is noted in utterance 2082 of Judy Bieber. It is due to a segment of trailing speech from the previous utterance that was not split correctly. Utterances with impersonated and animated speech fall among those with lower average WDP scores, but appear not to pose a problem for the forced alignment.

The manual verification of the bad alignments with "Too Long Insertion" is shown in Table 3.6. An important observation is the contrast that there are no word boundary errors (with most utterances scoring high above the threshold of -0.75) despite the presence of extra speech content in the form of chapter headers. Inspection of the whole subset of "Too Long Insertion" reveals that all of the members are utterances at the start of chapters. The contrast can be explained by way of the example of utterance 0001, spoken by Phil Chenevert, in Figure 3.4. Subfigure (a) illustrates how the alignment forces the start segment "SENT-START" to consume the whole chapter header, allowing the rest of the speech to be aligned properly to the body text, as in subfigure (b). This is the case for all of the utterances in the random selection.

From Table 3.7 it can be seen that the "Too Few Segments" criterion is mostly unnecessary as well. Most of the short utterances align well, with only animated speech and some other unknown factors causing word boundary errors. An example is given in Figure 3.5 of the difference in alignment quality of utterance 0611 between the animated speech of Phil Chenevert (subfigure (a)) and the neutral speech of Judy Bieber

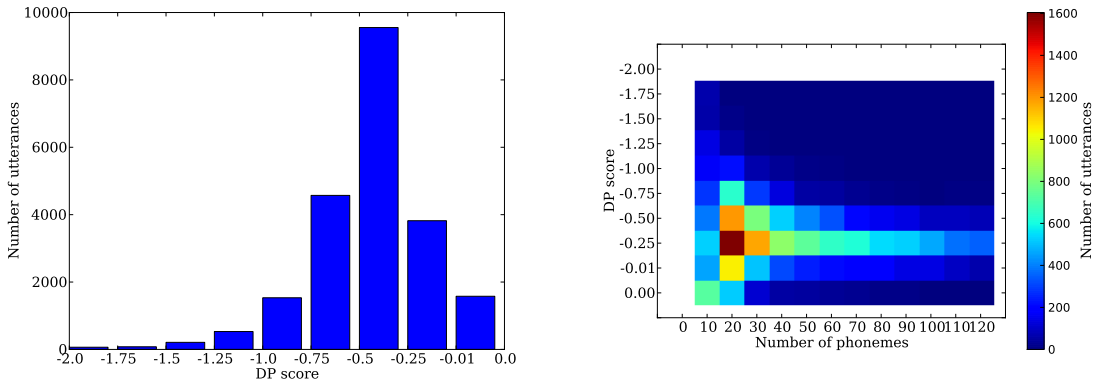
[2] An examiner notes that the phone-averaged score will always be the same over all insertions, since it is a property of the flat alignment matrix used in the DP technique. See Davel et al. (2012) for more details.

Table 3.1: Overall alignment statistics on the *Phil Chenevert* audiobooks

Book	**Overall Alignment Statistics**											
	All				**Good**				**Bad**			
	utts (audio)	**averages/utt**			**utts (audio)**	**averages/utt**			**utts (audio)**	**averages/utt**		
		W	**P**	**WDP**		**W**	**P**	**WDP**		**W**	**P**	**WDP**
oz01	2946 (04h06m)	14	45	-0.230	2444 (03h40m)	15	50	-0.183	502 (00h26m)	5	24	-0.461
oz03	3103 (04h11m)	13	46	-0.418	2469 (03h46m)	16	53	-0.364	634 (00h25m)	4	18	-0.629
oz04	3304 (04h23m)	13	46	-0.441	2610 (03h55m)	16	53	-0.381	694 (00h28m)	5	19	-0.668
oz05	3079 (04h20m)	14	47	-0.425	2455 (03h54m)	16	55	-0.374	624 (00h26m)	4	18	-0.624
oz07	4981 (06h38m)	12	42	-0.515	3739 (05h49m)	15	51	-0.441	1242 (00h48m)	4	16	-0.737
train	17413 (23h38m)	13	45	-0.420	13717 (21h04m)	16	52	-0.358	3696 (02h33m)	4	18	-0.649
oz06	4528 (05h56m)	13	44	-0.480	3532 (05h12m)	15	51	-0.418	996 (00h43m)	5	20	-0.700
all	21941 (29h34m)	13	45	-0.432	17249 (26h16m)	15	52	-0.370	4692 (03h16m)	4	19	-0.660

Table 3.2: Bad alignment statistics on the *Phil Chenevert* audiobooks

Book	**Bad Alignment Statistics**											
	Too Long Insertion				**Too Few Segments**				**Too Low WDP Score**			
	utts (audio)	**averages/utt**			**utts (audio)**	**averages/utt**			**utts (audio)**	**averages/utt**		
		W	**P**	**WDP**		**W**	**P**	**WDP**		**W**	**P**	**WDP**
oz01	31 (00h11m)	28	209	-0.373	345 (00h06m)	2	6	-0.283	126 (00h08m)	9	29	-0.972
oz03	28 (00h09m)	27	171	-0.463	411 (00h07m)	2	6	-0.490	195 (00h07m)	6	21	-0.945
oz04	27 (00h09m)	32	181	-0.506	407 (00h07m)	2	6	-0.503	260 (00h11m)	6	22	-0.943
oz05	30 (00h10m)	25	166	-0.474	439 (00h08m)	2	6	-0.527	155 (00h07m)	6	23	-0.925
oz07	35 (00h08m)	11	109	-0.605	675 (00h13m)	2	6	-0.598	532 (00h25m)	6	23	-0.923
train	151 (00h47m)	24	165	-0.487	2277 (00h41m)	2	6	-0.500	1268 (00h58m)	7	23	-0.936
oz06	37 (00h13m)	27	187	-0.490	548 (00h10m)	2	6	-0.533	411 (00h18m)	6	23	-0.941
all	188 (01h00m)	25	170	-0.488	2825 (00h51m)	2	6	-0.506	1679 (01h16m)	6	23	-0.937



(a) Over the WDP score only

(b) Over the WDP score and number of phonemes

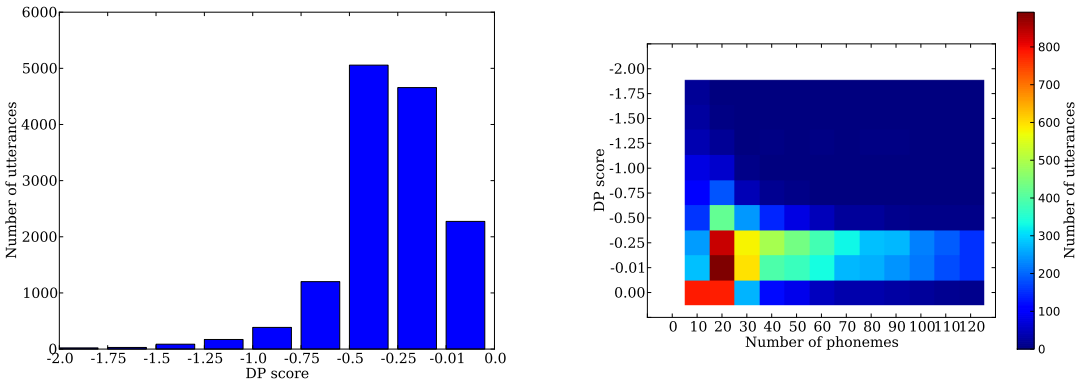
Figure 3.2: Average WDP score distribution (histograms) over all the *Phil Chenevert* audiobooks

Table 3.3: Overall alignment statistics on the *Judy Bieber* audiobooks

Book	**Overall Alignment Statistics**											
	All				**Good**				**Bad**			
	utts (audio)	**averages/utt**			**utts (audio)**	**averages/utt**			**utts (audio)**	**averages/utt**		
		W	**P**	**WDP**		**W**	**P**	**WDP**		**W**	**P**	**WDP**
oz03	3101 (04h07m)	13	46	-0.370	2524 (03h38m)	16	52	-0.324	577 (00h28m)	4	20	-0.574
oz04	3304 (04h15m)	12	45	-0.084	2796 (03h48m)	14	48	-0.069	508 (00h27m)	3	24	-0.165
oz10	2952 (05h02m)	18	62	-0.360	2534 (04h38m)	20	69	-0.316	418 (00h23m)	4	24	-0.629
train	9357 (13h24m)	14	51	-0.266	7854 (12h04m)	16	56	-0.231	1503 (01h18m)	4	22	-0.451
oz06	4526 (05h26m)	12	43	-0.323	3766 (04h50m)	14	48	-0.294	760 (00h36m)	4	21	-0.464
all	13883 (18h50m)	14	48	-0.284	11620 (16h54m)	16	54	-0.251	2263 (01h54m)	4	22	-0.456

Table 3.4: Bad alignment statistics on the *Judy Bieber* audiobooks

Book	**Bad Alignment Statistics**											
	Too Long Insertion				**Too Few Segments**				**Too Low WDP Score**			
	utts (audio)	**averages/utt**			**utts (audio)**	**averages/utt**			**utts (audio)**	**averages/utt**		
		W	**P**	**WDP**		**W**	**P**	**WDP**		**W**	**P**	**WDP**
oz03	20 (00h12m)	30	307	-0.512	423 (00h07m)	2	6	-0.451	134 (00h08m)	7	23	-0.971
oz04	31 (00h18m)	28	294	-0.227	462 (00h08m)	2	6	-0.138	15 (00h01m)	5	17	-0.877
oz10	22 (00h09m)	6	196	-0.584	276 (00h04m)	2	6	-0.452	120 (00h10m)	9	32	-1.046
train	73 (00h39m)	22	268	-0.413	1161 (00h19m)	2	6	-0.327	269 (00h19m)	7	27	-0.999
oz06	34 (00h17m)	26	263	-0.410	587 (00h10m)	2	6	-0.353	139 (00h08m)	7	25	-0.948
all	107 (00h56m)	23	266	-0.412	1748 (00h29m)	2	6	-0.336	408 (00h27m)	7	26	-0.982



(a) Over the WDP score only

(b) Over the WDP score and number of phonemes

Figure 3.3: Average WDP score distribution (histograms) over all the *Judy Bieber* audiobooks

(subfigure (b)). All the erroneous utterances but one have a low average WDP score. However, some of the utterances with no errors do too, for example utterance 1732.

Finally, Table 3.8 lists utterances that do not meet the aforementioned two criteria, but are flagged as below the -0.75 WDP threshold ("Too Low WDP Score"). Extraordinary speaking styles are present for many of the utterances, although they cause word boundary errors only in some cases. See Figure 3.6 for the example of utterance 1716 for both Phil Chenevert (subfigure (a)) and Judy Bieber (subfigure (b)). Chapter footers occur in some of the speech, but, like their header counterparts, are handled correctly by the forced alignment.

The quality of the good and supposedly bad alignments turn out to be very good. The average WDP threshold seems to be on the conservative side, but the problem is perhaps rather one of the appropriateness of the DP technique for prosodically rich audiobook speech. The magnitude of the average WDP score is necessarily dependent on the degree to which the actual phone behaviour (inter alia length) in the speech deviates from the mean phone behaviour modelled by the HMMs for free decoding. Therefore, speech segments with extraordinary prosody typically have worse WDP scores because the free decoding recognises (inter alia inserts) incorrect phones. Nevertheless, the recognition constraints imposed by the forced alignment is able to compensate for the deviations (even temporal ones) to produce robust phonetic alignments.

Revisiting the overall alignment statistics in Table 3.1 and Table 3.3, one last observation could provide a hint on improving the decoding accuracy and making the WDP scores more representative. Note the higher than usual WDP scores of the "oz01" entry for Phil Chenevert and the "oz04" entry for Judy Bieber. The speaker-specific seed models are trained on these audiobooks. It is thus evident that target data-specific acoustic models are more accurate at free decoding than speaker-specific models. Whether the former models would improve the quality of the already robust forced alignments remains future work. Another avenue that could be explored is more advanced acoustic models, such as the HMM and deep neural network (DNN) implementations in the toolkit Kaldi (Povey et al., 2011).

3.4 Conclusion

This chapter advocated the use of audiobook text and speech to model discourse, information structure and affect because its narrative is more suitable to evoke these phenomena than the traditional TTS corpus. The Oz series of audiobooks by the writer L. Frank Baum and the speakers Phil Chenevert and Judy Bieber were chosen for subsequent experimentation. These children's stories have grammatical and moral simplifications that should aid in the NLP development and performance of a model of affect, discussed in the next chapter.

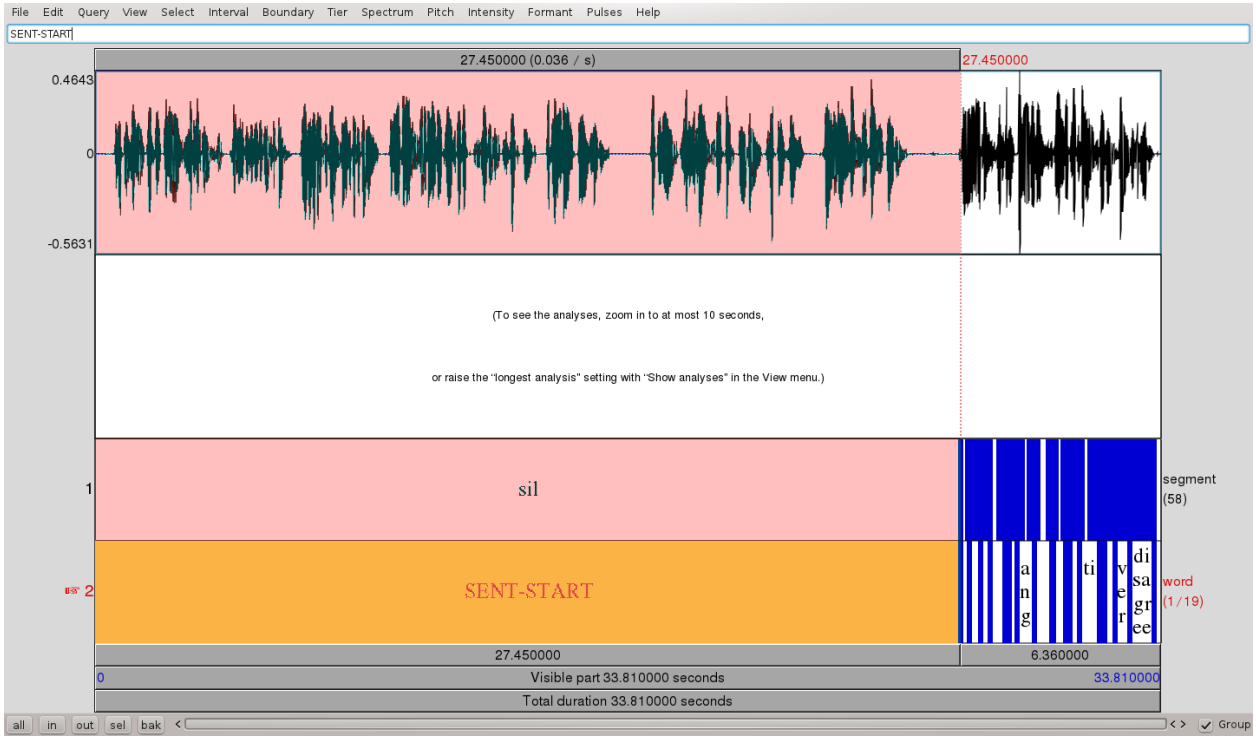
The alignments of the audiobooks are particularly robust, such that the quality control was perhaps too conservative in the amount of data that it kept. The DP technique used to score the alignments might not be suitable for prosodically varying audiobook speech. It is based on the closeness between the forced alignment and free decoding of an utterance, but the latter struggles to recognise prosodically rich phones, whereas the former can compensate for them. This produces artificially low scores. Be that as it may, the "good" data of the two speakers are substantial enough for analysis and will be used as is in the prosodic investigation of Chapter 5.

Table 3.5: Manual verification of randomly selected *good* alignments

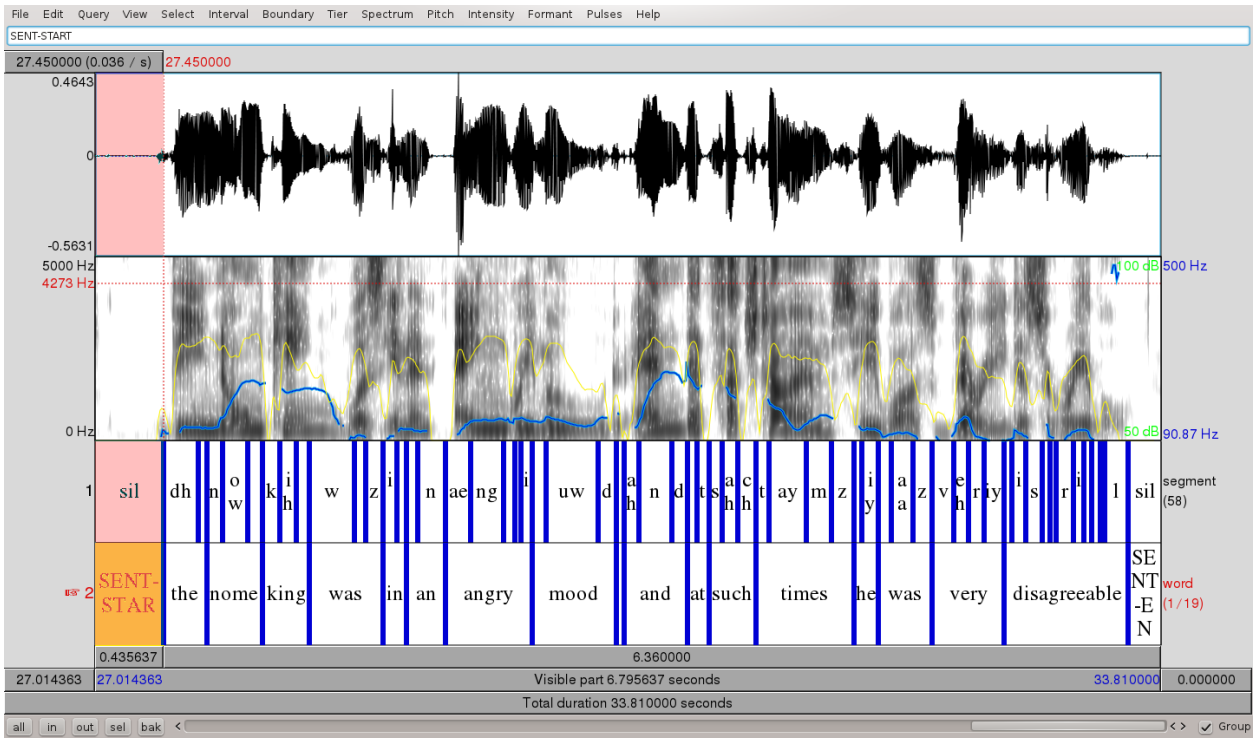
Utterance		Phil Chenevert			Judy Bieber		
No	**Text**	**WDP**	**E**	**Meta**	**WDP**	**E**	**Meta**
0311	And whenever you signal me to transport you to this safe place, where you are always welcome, I know you are in danger or in trouble."	-0.222	0		-0.168	0	
0410	replied Colonel Crinkle, a dapper-looking Nome, as he stepped forward to salute his monarch.	-0.576	0		-0.247	0	
0479	replied Guph, calmly, and he blew a wreath of smoke that curled around the King's nose and made him sneeze.	-0.269	0		-0.304	0	1× sub
1253	You see, the chicken had eaten an Elocution Pill."	-0.617	0		-0.257	0	
1413	Upon a table were paints and brushes, while several pair of scissors, of different sizes, were lying about.	-0.209	0		-0.329	0	
1940	inquired the King, suspiciously, for he knew how greedy the Growleywogs were.	-0.603	0		-0.333	0	
2031	Do what you will with the girls for all I care.	-0.423	0		-0.367	0	
2082	gasped Aunt Em.	-0.267	0		-0.653	1	
2189	"But there is nothing in the kettle,"	-0.671	0	person	-0.650	0	animate
2388	"And you are our prisoners,"	-0.522	0	animate	-0.388	0	animate
2501	It is only necessary to kill a person once to make him dead; but I do not see that it is necessary to kill this little girl at all."	-0.257	0		-0.248	0	
2724	Toto ate some, too, while Billina picked up the crumbs.	-0.493	0		-0.403	0	
2841	She soon decided to make Uncle Henry the Keeper of the Jewels, for some one really was needed to count and look after the bins and barrels of emeralds, diamonds, rubies and other precious stones that were in the Royal Storehouses.	-0.342	0		-0.179	0	
3458	replied the Wizard;	-0.083	0		0.000	0	
3685	"There's another 'if',"	-0.500	0	animate	-0.500	0	
3905	"What can be done?"	-0.467	0	animate	-0.708	0	
3948	You may imagine how big this ear of corn was when I tell you that a single gold kernel formed a window, swinging outward upon hinges, while a row of four kernels opened to make the front entrance.	-0.434	0	1× sub	-0.375	0	
3977	It is surely time enough to be sad when our country is despoiled and our people made slaves.	-0.493	0		-0.267	0	
4363	The sun came up and sent its flood of silver rays to light the faces of the invaders.	-0.517	0		-0.343	0	
4376	The only question remaining to solve was how to get rid of this horde of intruders.	-0.733	0		-0.239	0	

Table 3.6: Manual verification of randomly selected *bad* alignments ("Too Long Insertion")

Utterance		Phil Chenevert			Judy Bieber		
No	**Text**	**WDP**	**E**	**Meta**	**WDP**	**E**	**Meta**
0001	The Nome King was in an angry mood, and at such times he was very disagreeable.	-0.364	0	ch start	-0.853	0	ch start
0164	Dorothy Gale lived on a farm in Kansas, with her Aunt Em and her Uncle Henry.	-0.354	0	ch start	-0.346	0	ch start
0381	The reason most people are bad is because they do not try to be good.	-0.447	0	ch start	-0.403	0	ch start
0533	When the people of the Emerald City heard that Dorothy had returned to them every one was eager to see her, for the little girl was a general favorite in the Land of Oz.	-0.312	0	ch start	-0.127	0	ch start
0674	The new General of the Nome King's army knew perfectly well that to fail in his plans meant death for him.	-0.510	0	ch start	-0.515	0	ch start
0740	"These are your rooms,"	-0.760	0	ch start	-0.814	0	ch start
1055	It did not take Dorothy long to establish herself in her new home, for she knew the people and the manners and customs of the Emerald City just as well as she knew the old Kansas farm.	-0.389	0	ch start	-0.351	0	ch start
2225	It was a beautiful evening, so they drew their camp chairs in a circle before one of the tents and began to tell stories to amuse themselves and pass away the time before they went to bed.	-0.483	0	ch start	-0.386	0	ch start
2422	There must have been from six to eight dozen spoons in the Brigade, and they marched away in the shape of a hollow square, with Dorothy, Billina and Toto in the center of the square.	-0.433	0	ch start	-0.345	0	ch start
2614	Wandering through the woods, without knowing where you are going or what adventure you are about to meet next, is not as pleasant as one might think.	-0.435	0	ch start	-0.284	0	ch start
2874	Dorothy left Bunbury the same way she had entered it and when they were in the forest again she said to Billina:	-0.340	0	ch start	-0.376	0	ch start
3027	A line of rabbit soldiers was drawn up before the palace entrance, and they wore green and gold uniforms with high shakos upon their heads and held tiny spears in their hands.	-0.290	0	ch start	-0.337	0	ch start
3240	Just then a rabbit band of nearly fifty pieces marched in, playing upon golden instruments and dressed in neat uniforms.	-0.511	0	ch start	-0.449	0	ch start
3385	When they came to the signpost, there, to their joy, were the tents of the Wizard pitched beside the path and the kettle bubbling merrily over the fire.	-0.311	0	ch start	-0.409	0	ch start
3563	They were soon among the pretty hills and valleys again, and the Sawhorse sped up hill and down at a fast and easy pace, the roads being hard and smooth.	-0.421	0	ch start	-0.408	0	ch start
3911	This amazing news had saddened every heart and all were now anxious to return to the Emerald City and share Ozma's fate.	-0.392	0	ch start	-0.296	0	ch start
4239	The Nome King and his terrible allies sat at the banquet table until midnight.	-0.729	0	ch start	-0.545	0	ch start
4307	The Scarecrow had no need to sleep; neither had the Tin Woodman or Tiktok or Jack Pumpkinhead.	-0.414	0	ch start	-0.378	0	ch start
4412	"That was better than fighting,"	-0.763	0	ch start	-0.195	0	ch start
4520	The writer of these Oz stories has received a little note from Princess Dorothy of Oz which, for a time, has made him feel rather disconcerted.	-0.340	0	ch start	-0.438	0	ch start

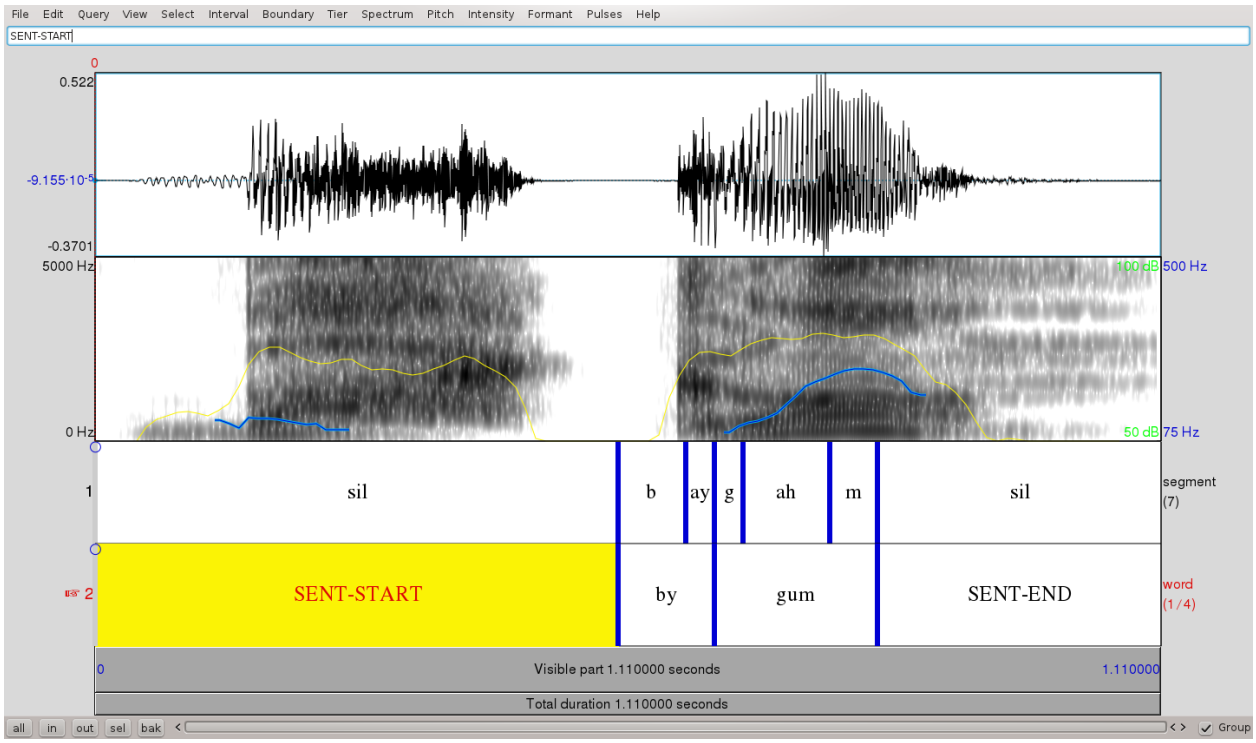


(a) Start segment consumes chapter header

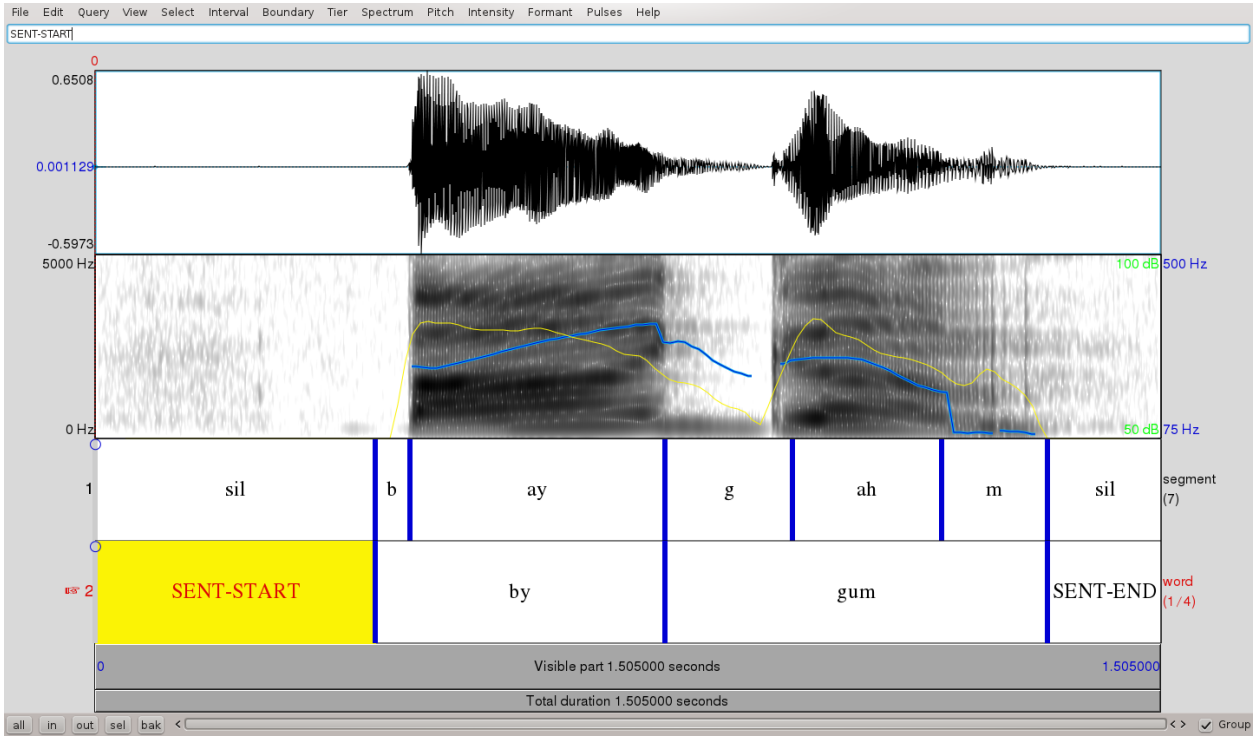


(b) Subsequent segments align correctly

Figure 3.4: Waveform analysis of the forced alignment of utterance 0001 ("Too Long Insertion") for Phil Chenevert

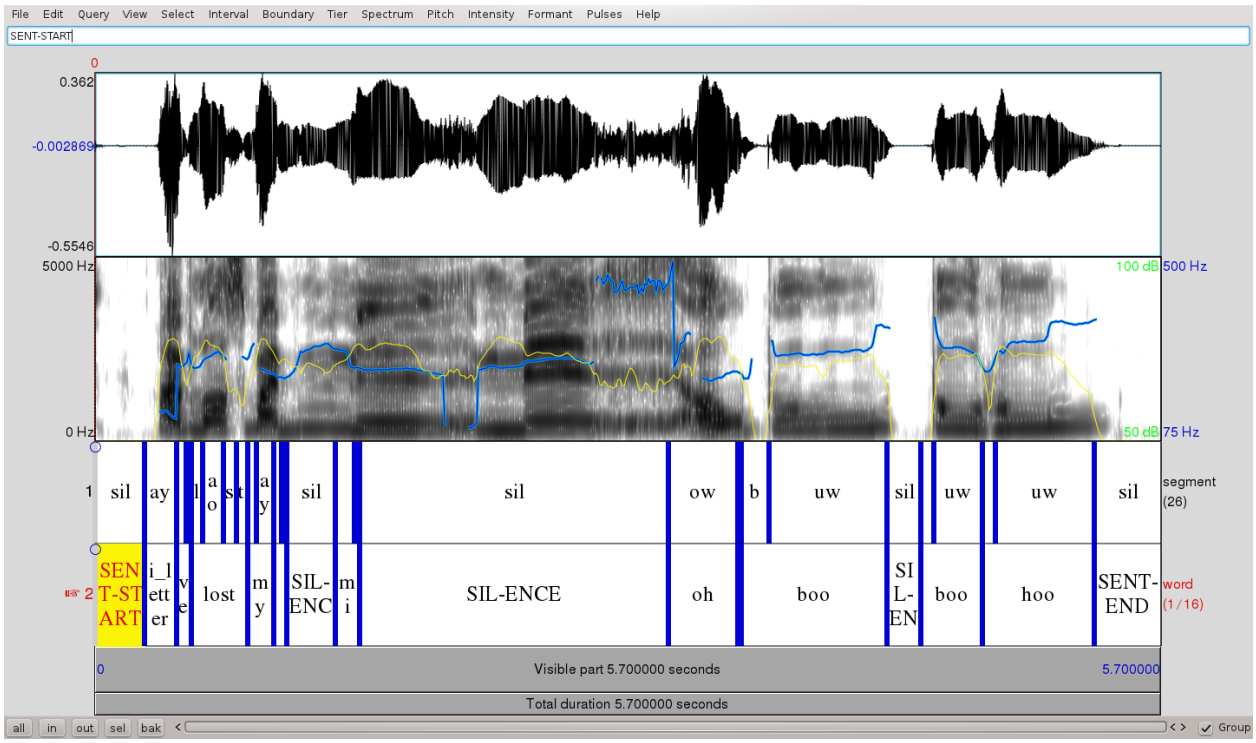


(a) Phil Chenevert

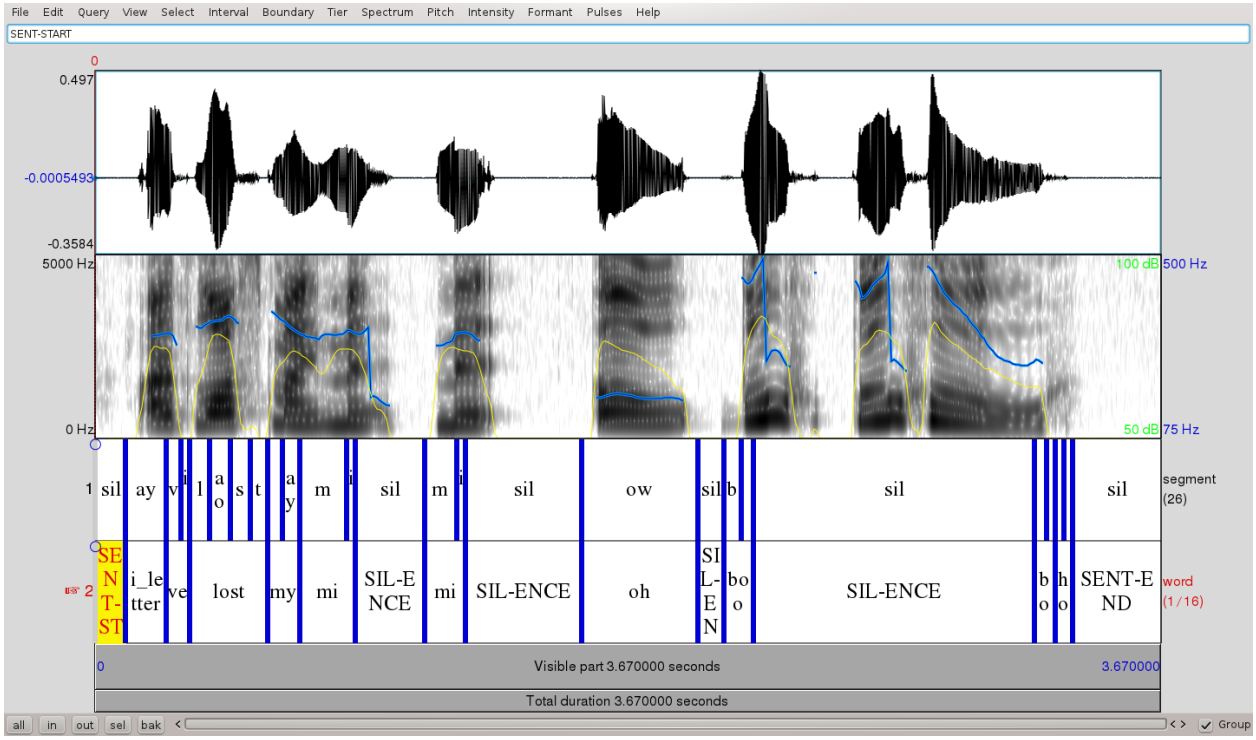


(b) Judy Bieber

Figure 3.5: Waveform analysis of the forced alignment of utterance 0611 ("Too Few Segments")



(a) Phil Chenevert



(b) Judy Bieber

Figure 3.6: Waveform analysis of the forced alignment of utterance 1716 (“Too Low WDP Score”)

Table 3.7: Manual verification of randomly selected *bad* alignments ("Too Few Segments")

Utterance		Phil Chenevert			Judy Bieber		
No	**Text**	**WDP**	**E**	**Meta**	**WDP**	**E**	**Meta**
0101	What's that?"	-0.250	0	animate	-0.167	0	
0611	"By gum!"	-1.292	3	animate	-0.709	0	
1302	said Dorothy.	-0.250	0		-0.250	0	
1391	"Here we are!"	-0.219	0		-0.125	0	
1732	"Boo-hoo!	-1.625	0		-1.583	0	
2094	said Dorothy,	-0.250	0		0.000	0	
2147	said he,	0.000	0		0.000	0	
2385	said one.	0.000	0		-0.167	0	
2471	she replied.	0.000	0		-0.167	0	
2595	"Discharged!"	-1.000	0		-0.833	0	animate
2698	"So do I, Pop,"	-0.667	0		-1.458	0	
2810	"Oh, Toto!	-0.444	0		0.000	0	
2978	she said,	0.000	0		0.000	1	
3166	she asked.	0.000	0		0.000	0	
3491	"Now,"	-1.167	1		0.000	0	
3524	"Not here,"	-0.438	0		-0.334	0	
3642	"Help!"	-0.875	1	animate	-0.300	0	
4023	he said.	0.000	0		-0.167	0	
4148	I've been there,"	-0.979	0		-0.312	0	
4264	"Good!"	-1.667	1	animate	0.000	0	

Table 3.8: Manual verification of randomly selected *bad* alignments ("Too Low WDP Score")

Utterance		Phil Chenevert			Judy Bieber		
No	**Text**	**WDP**	**E**	**Meta**	**WDP**	**E**	**Meta**
0857	Come in the house and I'll show it to you.	-0.890	1		-0.826	0	
1292	Then let's go somewhere else,"	-0.799	0	person	-0.933	0	
1716	I've lost my mi--mi--Oh, boo, boo-hoo!"	-0.856	4	person	-0.774	6	
1727	"I've lost my mi--mi--mittens!"	-1.042	4	person	-0.787	5	
1749	Where do you live?"	-1.538	0	question	-1.188	0	question
1870	"there used to be a picture puzzle craze in Kansas, and so I've had some 'sperience matching puzzles.	-0.825	0	person	-0.754	0	
1913	"For my part, I'm glad we visited the Fuddles."	-1.298	0	ch end	-1.427	0	ch end
1980	"When will they come?"	-0.775	0	person	-0.865	0	person
2034	promised the King, and hurried away to inspect the work and see that the Nomes kept busy.	-0.977	0	ch end	-1.466	0	
2049	"There isn't any path,"	-0.896	0		-0.854	0	
2107	"Perhaps we're lost,"	-1.292	0		-0.917	0	
2224	For, sure enough, when she looked at the dishes they had a moment before left upon the table, she found them all washed and dried and piled up into neat stacks.	-0.932	0	ch end	-1.303	0	ch end
2296	"The soft-shell crab is correct,"	-0.895	0		-0.839	0	
2362	"I didn't try to remember,"	-0.971	0	person	-0.774	0	
2464	Get off from me!"	-1.292	2	person	-1.125	0	animate
3069	"Yes, your Majesty,"	-0.764	0		-0.868	0	animate
3401	"I'm not hungry.	-0.812	0		-0.792	0	
3515	cried Aunt Em, impatiently;	-0.864	0		-0.840	0	
3653	he explained;	-0.875	0		-1.125	0	
4179	"Why do they call it the Forbidden Fountain?"	-0.782	0	question	-0.845	0	question

Chapter 4

A Discourse Model of Affect

4.1 Introduction

Chapter 3 established the audiobook as a source of discourse-level linguistic and prosodic phenomena towards modelling affect. However, Chapter 2 also concluded that a model of affect should be grounded in emotion theory. In particular, this chapter will describe the OCC model of Ortony et al. (1988) and its implementation by Shaikh et al. (2009a), and discuss its suitability for the audiobook domain. A new model will be introduced that should address the shortcomings of the aforementioned implementation, and its performance will be tested on audiobook text. The chapter will conclude with closing remarks on the new model.

4.2 The OCC Model

The OCC model of Ortony et al. (1988) is a cognitive model of emotion that is a point of departure for this study because it takes a step back from the surface level of emotional expressions and rather identifies the underlying factors that contribute towards them. It has been implemented in the personality design of artificial agents (Bartneck, 2002; Steunebrink, 2010) and, relevant to this study, in affect detection from text (Shaikh et al., 2009a).

4.2.1 Specification

Expounding from Section 2.4.4, the OCC model appraises[1] 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. For example, the event of:

```
I shot the sheriff.
```
(4.1)

[1] An appraisal is something perceived and valued.

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).

Figure 4.1 depicts a tree diagram of the OCC model from this focus-of-attention view. An appraisal of a particular emotion is a path along the tree that starts at the root, which defines a primitive positive or negative reaction, and ends at one of the leaves, which defines a more complex emotion, depending on the focus. For example, the emotion of *pride* is a positive reaction to (being approving of) one's own praiseworthy action. The OCC model also distinguishes emotions based on the actions of or consequences for others. For example, the emotion of *pity* is a negative reaction to (being displeased about) an event presumed to be undesirable for someone else.

The desirability or undesirability of the consequences of an event for "self" is implicitly specified in its positive or negative reaction, since "self" will not contradict its own goals. However, the desirability or undesirability for "other" needs to be specified explicitly, since this might be in opposition to the goals of "self". The model also caters for prospective events. The emotions of *hope* and *fear* are leaves when the prospect of an event is unconfirmed. When the latter is confirmed or disconfirmed, the emotions of *satisfaction*, *relief*, et cetera become the leaves. The emotions of *gratification*, *gratitude*, et cetera are compound leaves when the focus lies on both the action of the agent ("self" or "other") responsible for the event and the consequences of the event for "self".

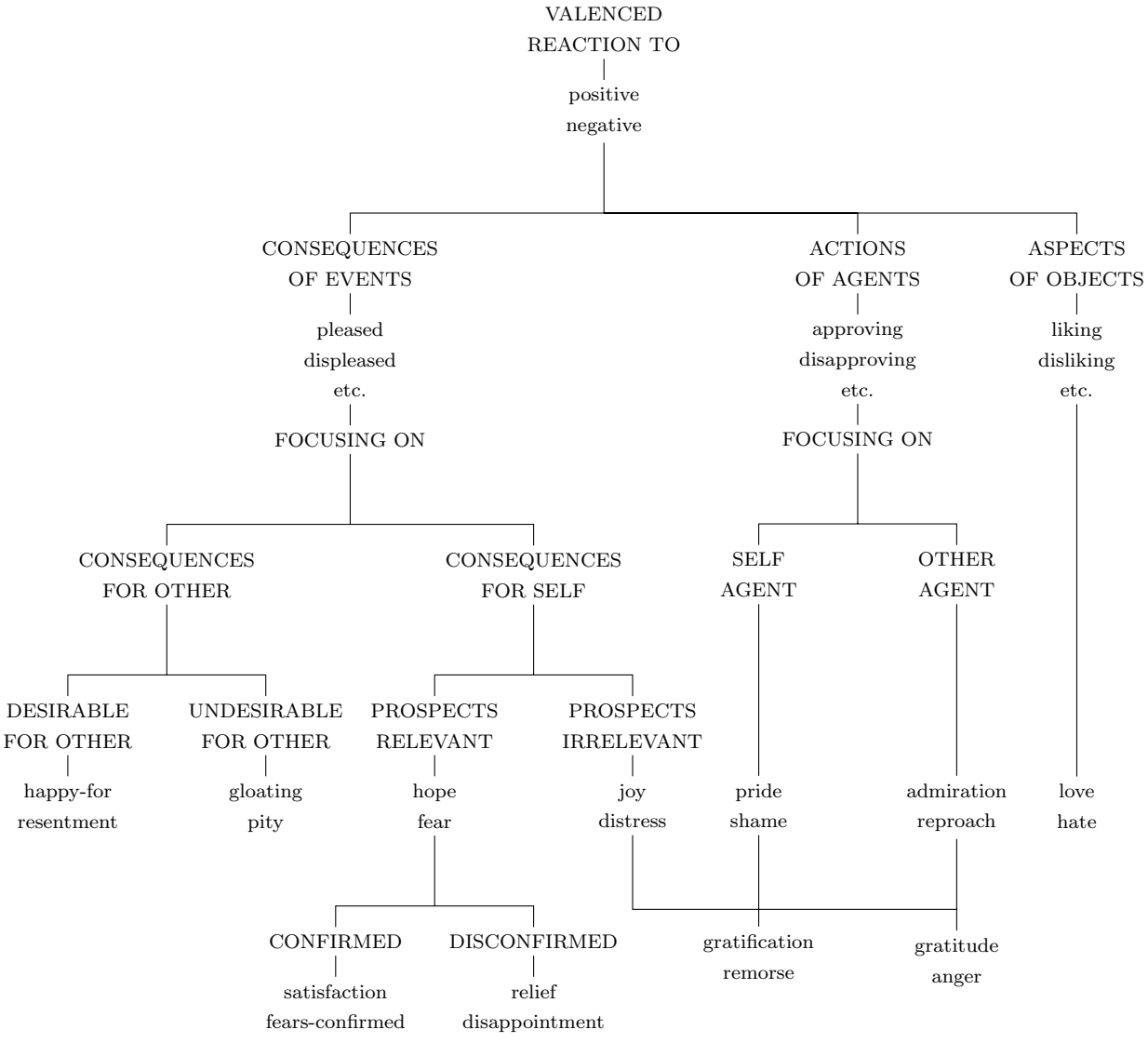


Figure 4.1: The OCC model (focus-of-attention view)

Finally, the OCC model specifies variables that affect the intensity of the appraised emotions. For example, the emotion of *fear*, which is a negative reaction to (being displeased about) the prospect of an undesirable event, has the following variables affecting its intensity:

1. The degree to which the event is undesirable
2. The likelihood of the event

4.2.2 Strengths and Weaknesses

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. This is exactly what Shaikh et al. (2009a) have used in their implementation to good effect.

The difficulty lies with the cognitive antecedents. Whereas they can be preprogrammed explicitly into the personalities of artificial agents (Bartneck, 2002; Steunebrink, 2010)—making these implementations of the OCC model successful thus far—they are much harder to infer implicitly from text. Shaikh et al. (2009a) make the simplifying, but pragmatic, assumption that all the antecedents are subsumed by the *lexical affinity* of a word, that is its positive or negative sentiment ("good" or "bad"). Their sentiment analysis system, called SenseNet (Shaikh et al., 2008b), assigns a prior valence on a numerical scale to individual nouns, verbs, adjectives and adverbs that is based upon the net positive sentiment count of all the senses of the particular word found in WordNet (extended by ConceptNet). Named entities are scored with a general opinion automatically mined from the internet. The valences of noun phrases (containing nouns and adjectives) and verb phrases (containing verbs and adverbs) are composed from the individual word valences by boolean conjunction (AND). Any occurrence of negation (by negators such as `not`) is handled by boolean negation (NOT).

During the course of its assignment of the various OCC variables, the implementation specifies certain composition rules for the individual semantic role valences, which may be likened to ascribing to a particular belief system. For example, the valence for the reaction to an OCC event is composed from the valences of the predicate verb phrase and the PATIENT noun phrase, according to the objective belief system of *justice* (fairness), hence depending on the *deservingness* of the entity in the PATIENT role:

- If a negatively valenced entity (semantic PATIENT) undergoes a negatively valenced action (semantic predicate), then the OCC event receives a positive valence.
- If a positively valenced entity undergoes a negatively valenced action, then the event receives a negative valence.
- If a negatively valenced entity undergoes a positively valenced action, then the event receives a negative valence.
- If a positively valenced entity undergoes a positively valenced action, then the event receives a positive valence.

As noted in the previous section, the desirability of the consequences of the event for "self" is implied in the aforementioned valenced reaction. The desirability of the consequences for "other", however, is obtained by composing the valences of the AGENT noun phrase and the OCC event in a subjective, or relativist, manner contrary to justice (reciprocation):

- If a negatively valenced entity (semantic AGENT) performs a negatively valenced OCC event (semantic predicate and PATIENT), then the consequence for "other" is desirable.
- If a negatively valenced entity performs a positively valenced event, then the consequence is undesirable.
- If a positively valenced entity performs a negatively valenced event, then the consequence is undesirable.
- If a positively valenced entity performs a positively valenced event, then the consequence is desirable.

The implementation furthermore resolves prospective emotions effectively by identifying the tense of the verbs in the sentence. Present and future tense indicates unconfirmed prospects, past tense indicates confirmed prospects and negated past tense indicates disconfirmed prospects. The intensity variables for a particular emotion are set according to the positions of the constituent valences on the numerical scale. The overall accuracy of the system is 80.5% on a 200 sentence test set when the 22 OCC emotions are collapsed onto the 6 basic emotions of joy, sadness, fear, anger, disgust and surprise (for comparison to related work).

The implementation of Shaikh et al. (2009a) has a few weaknesses generally, and specifically with regards to the precondition of an audiobook narrative domain:

- The valence composition rules are contradictory, assuming both an objective and subjective belief system simultaneously.
- The valences of named entities are based on general, real-world opinion, whereas the characters in a narrative are fictional and need to have their valences induced from the text itself.
- Furthermore, the valence assignment is static, whereas the characters can change status as the discourse progresses.
- The range of the OCC focus of attention is insufficient when the environmental factors are mapped to semantic predicates, AGENTs and PATIENTs. It does not allow a focus on the consequences for the PATIENT.
- Although the OCC notion of "self" can be analogous to the narrator in the case of indirect speech narrative, it does not cater for the situation in direct speech dialogue where the narrator can appraise emotions vicariously on behalf of the interlocutors.

4.3 e-*motif*

A new model, named e-*motif*, will now be put forth to address the weaknesses of the OCC model implementation of Shaikh et al. (2009a). 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*.

4.3.1 Judgment

It was noted in the previous section that 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. This Shaikh et al. (2009a) did, and e-*motif* follows suit: it is also 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, 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. If e-*motif* is restricted to the cognitive feature context, it 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, or good always deserves bad and bad always deserves good, et cetera (In Section 4.3.4, however, it will be argued that the social feature of power provides more flexibility). The number of possible affective states produced by e-*motif* is $2^3 = 8$, as illustrated in Table 4.1. 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.
```
(4.2)

Table 4.1: 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.*

The compatibility of e-*motif* with the established OCC model rationale can be demonstrated if an objective belief system of justice is assumed in the appraisal of the OCC emotions, and if the following simplifications to the OCC model are made (compare Figure 4.2 with Figure 4.1):

- An event always constitutes the action of an AGENT on a PATIENT. This means that gratification and remorse collapse onto pride and shame, and gratitude and anger onto admiration and reproach.
- No distinction is made between the OCC "other" and "self"; "self" is simply another participant (AGENT or PATIENT) in the event. This means that the consequences for other apply to self, and joy and distress collapse onto happy-for and resentment. Furthermore, (the new) pride and shame collapse onto (the new) admiration and reproach.
- No distinction is made between OCC objects and agents; an object is simply another AGENT or PATIENT (however, the *liking* and *disliking* factors are featured implicitly in the judicial valences of the AGENT and PATIENT). This means that love and hate collapse onto (the very new) admiration and reproach.
- Temporal semantics are ignored for this example. This means that the prospective emotions hope and fear and their derivatives satisfaction and fears-confirmed, and relief and disappointment fall away.

Note the following truth tables for the three focus areas of the simplified OCC model, namely the action of the AGENT and the consequences of the event for the AGENT and the PATIENT. Table 4.2 appraises the

action of the AGENT according to justice: a bad/good deed to a bad/good PATIENT is overall bad/good depending on the *deservingness* of the PATIENT. Table 4.3 shows the judicial appraisal of the consequences ("conseq") for the PATIENT: a bad/good deed to a bad/good PATIENT is a trivial one-to-one bad/good consequence for the PATIENT, and overall bad/good depending on the deservingness of the PATIENT. Table 4.4 depicts the consequence for the AGENT as a judicial *reciprocation* of the overall appraisal of his action on the PATIENT, and overall bad/good depending on the deservingness of the AGENT. Finally, Table 4.5 summarises the emotions for each focus area.

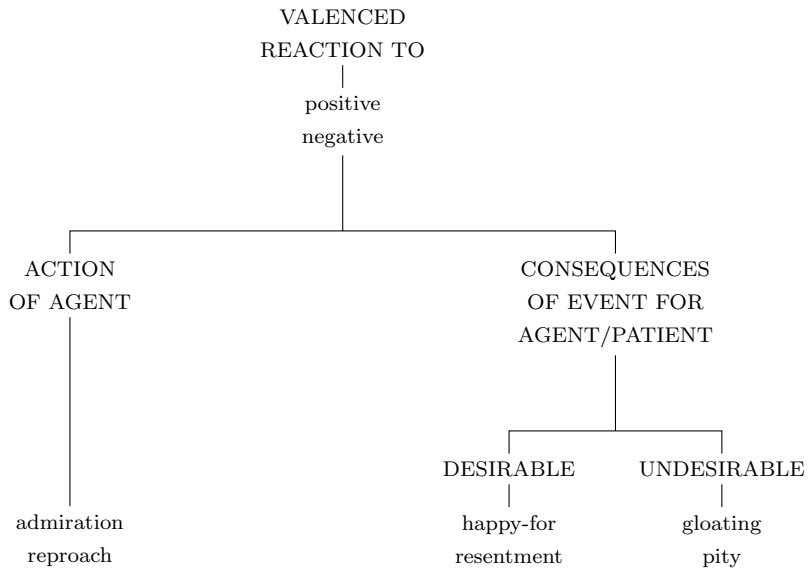


Figure 4.2: Simplified OCC model for e-*motif*

4.3.1.1 Implementation

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. The following list describes the choices around valence defaults, modification and negation:

- 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 (Esuli and Sebastiani, 2006) assigns a positive or negative sentiment score to each WordNet (Fellbaum, 1999) entry. If no entry is available, a good valence is assigned. e-*motif* follows the methodology of Shaikh et al. (2009a) 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 Shaikh et al. (2009a), 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).

Table 4.2: Truth table for the appraisal of the action of the AGENT

			Appraisal (dependent on v**P**)			
A	v	**P**	**Reaction**	**Rationale**	**Emotion**	**Rationale**
0	0	0	positive	bad deed to bad P is good	admiration	positive reaction to action of A
0	0	1	negative	bad deed to good P is bad	reproach	negative reaction to action of A
0	1	0	negative	good deed to bad P is bad	reproach	negative reaction to action of A
0	1	1	positive	good deed to good P is good	admiration	positive reaction to action of A
1	0	0	positive	bad deed to bad P is good	admiration	positive reaction to action of A
1	0	1	negative	bad deed to good P is bad	reproach	negative reaction to action of A
1	1	0	negative	good deed to bad P is bad	reproach	negative reaction to action of A
1	1	1	positive	good deed to good P is good	admiration	positive reaction to action of A

Table 4.3: Truth table for the appraisal of the consequences for the PATIENT

			Appraisal (dependent on v**P**)			
A	v	**P**	**Reaction**	**Rationale**	**Emotion**	**Rationale**
0	0	0	positive	bad conseq for bad P is good	gloating	positive reaction to undesirable conseq for P
0	0	1	negative	bad conseq for good P is bad	pity	negative reaction to undesirable conseq for P
0	1	0	negative	good conseq for bad P is bad	resentment	negative reaction to desirable conseq for P
0	1	1	positive	good conseq for good P is good	happy-for	positive reaction to desirable conseq for P
1	0	0	positive	bad conseq for bad P is good	gloating	positive reaction to undesirable conseq for P
1	0	1	negative	bad conseq for good P is bad	pity	negative reaction to undesirable conseq for P
1	1	0	negative	good conseq for bad P is bad	resentment	negative reaction to desirable conseq for P
1	1	1	positive	good conseq for good P is good	happy-for	positive reaction to desirable conseq for P

Table 4.4: Truth table for the appraisal of the consequences for the AGENT

			Appraisal (dependent on **A**v**P**)			
A	v	**P**	**Reaction**	**Rationale**	**Emotion**	**Rationale**
0	0	0	negative	good conseq for bad A is bad	resentment	negative reaction to desirable conseq for A
0	0	1	positive	bad conseq for bad A is good	gloating	positive reaction to undesirable conseq for A
0	1	0	positive	bad conseq for bad A is good	gloating	positive reaction to undesirable conseq for A
0	1	1	negative	good conseq for bad A is bad	resentment	negative reaction to desirable conseq for A
1	0	0	positive	good conseq for good A is good	happy-for	positive reaction to desirable conseq for A
1	0	1	negative	bad conseq for good A is bad	pity	negative reaction to undesirable conseq for A
1	1	0	negative	bad conseq for good A is bad	pity	negative reaction to undesirable conseq for A
1	1	1	positive	good conseq for good A is good	happy-for	positive reaction to desirable conseq for A

Table 4.5: Truth table that summarises the emotions for each focus area in the appraisal of an event

			Emotion		
A	v	**P**	**Action of A**	**Consequence for P**	**Consequence for A**
0	0	0	admiration	gloating	resentment
0	0	1	reproach	pity	gloating
0	1	0	reproach	resentment	gloating
0	1	1	admiration	happy-for	resentment
1	0	0	admiration	gloating	happy-for
1	0	1	reproach	pity	pity
1	1	0	reproach	resentment	pity
1	1	1	admiration	happy-for	happy-for

Negation is applied straightforwardly after modification by boolean negation (NOT):

$$\begin{aligned}
\texttt{nasty}^0 \wedge \texttt{criminal}^0 &= \texttt{(nasty criminal)}^0 \\
\texttt{nasty}^0 \wedge \texttt{policeman}^1 &= \texttt{(nasty policeman)}^0 \\
\texttt{nice}^1 \wedge \texttt{criminal}^0 &= \texttt{(nice criminal)}^0 \\
\texttt{nice}^1 \wedge \texttt{policeman}^1 &= \texttt{(nice policeman)}^1 \\
\neg\texttt{(nasty criminal)}^0 &= \texttt{(not nasty criminal)}^1
\end{aligned} \tag{4.3}$$

- 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:

$$\begin{aligned}
\texttt{to kill}^0 \wedge \texttt{savagely}^0 &= \texttt{(to kill savagely)}^0 \\
\texttt{to kill}^0 \wedge \texttt{gently}^1 &= \texttt{(to kill gently)}^0 \\
\texttt{to save}^1 \wedge \texttt{savagely}^0 &= \texttt{(to save savagely)}^0 \\
\texttt{to save}^1 \wedge \texttt{gently}^1 &= \texttt{(to save gently)}^1 \\
\neg\texttt{(to kill savagely)}^0 &= \texttt{(not to kill savagely)}^1
\end{aligned} \tag{4.4}$$

- Absent semantic roles receive a good valence.

Initially, the valences of the discourse entities and actions are based on the general, real-world opinion of SentiWordNet. e-*motif* allows the valences to change to accommodate the fictional domain by induction from the text:

- 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.
- Discourse entities can receive new valences as in the following examples:

$$\begin{aligned}
\texttt{Jack}^0\ \texttt{is a criminal}^0 &\rightarrow \texttt{Jack}^0 \\
\texttt{Jack}^0\ \texttt{is a policeman}^1 &\rightarrow \texttt{Jack}^1 \\
\texttt{Jill}^1\ \texttt{is bad}^0 &\rightarrow \texttt{Jill}^0 \\
\texttt{Jill}^1\ \texttt{is good}^1 &\rightarrow \texttt{Jill}^1
\end{aligned} \tag{4.5}$$

- Discourse actions follow suit:

$$\begin{aligned}
\texttt{to kill}^0\ \texttt{is a crime}^0 &\rightarrow \texttt{to kill}^0 \\
\texttt{to kill}^0\ \texttt{is a virtue}^1 &\rightarrow \texttt{to kill}^1 \\
\texttt{to save}^1\ \texttt{is wrong}^0 &\rightarrow \texttt{to save}^0 \\
\texttt{to save}^1\ \texttt{is right}^1 &\rightarrow \texttt{to save}^1
\end{aligned} \tag{4.6}$$

- Pragmatically, the copular induction is forced to hold across the storyboard of characters. In other words, for the sake of the narrator's judgment, in direct speech dialogue, the interlocutors are assumed

always to tell the truth about one another and never to contradict one another when making copular assertive statements. This assumption is deemed necessary since situations may arise where the valences of the characters are not introduced in the preceding indirect speech narrative. At least, children's stories should be amenable to such moral simplifications more often than other genres. In fact, one such example already occurs in the beginning of the second chapter of the first book in the Oz series:

```
"But who was she?" asked Dorothy.
"She was the Wicked Witch of the East, as I said," answered the little woman.
```
(4.7)

4.3.2 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 (v) 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 4.6 shows how the information status values can combine to form proper theme and/or rheme phrase sequences according to Steedman (1991, 2000, 2007). *Importantly, this cognitive feature of focus in e-motif subsumes the prosodic effects of information structure under those of affect.*

Table 4.6: Truth table for the focus areas in e-*motif*

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

4.3.2.1 Implementation

The "current discourse" is defined as the current $\mathbf{A}v\mathbf{P}$ clause and the "immediately preceding discourse" as the previous $\mathbf{A}v\mathbf{P}$ 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.

4.3.3 Tense

As in Shaikh et al. (2009a), 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.

4.3.3.1 Implementation

Tense is captured in the POS tags of the verbs as output by Stanford CoreNLP.

4.3.4 Power

Section 2.4.4 noted the social factor of *power* that 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 4.7. The discourse context for the examples in the table is again (4.2).

Table 4.7: Possible combinations of SPEAKER-LISTENER power

S	**L**	**Gloss**	**Example**
0	0	bad S speaking to bad L	*The criminal said to the other criminal, ...*
0	1	bad S speaking to good L	*The criminal shouted at the policeman, ...*
1	0	good S speaking to bad L	*The policeman reprimanded the criminal, ...*
1	1	good S speaking to good L	*The policeman answered the other 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.

4.3.4.1 Implementation

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 still no SPEAKER is found for the current speech report, look in the dialogue turn history and assign the previous LISTENER.
- If still no LISTENER is found for the current speech report, look in the indirect speech narrative for a discourse 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.

4.3.5 Interaction

This feature models the social responses of the characters in their direct speech *interaction* with one another and the environment. To reiterate from Section 2.4.4:

- *Adaptation* captures the adjustment of a character in response to the environment.
- *Coordination* captures the reaction of one character in response to the emotional expressions of another.
- *Regulation* captures the reaction of a character based on his understanding of his own emotional state and relationship with the environment.

These definitions appear to overlap somewhat with that of the OCC model, but the intent here is to add another dimension to direct speech discourse to distinguish it from indirect speech narrative discourse. The following description of their implementation will clarify the matter.

4.3.5.1 Implementation

e-*motif* straightforwardly assigns the feature values to the direct speech segments, based on their interaction with one another and the indirect speech narrative:

- Adaptation 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 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 is set for each non-initial clause in the direct speech monologue sequence of a character in his dialogue turn.

4.3.6 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.

4.3.6.1 Implementation

The presence of a question mark in a sentence-final position very simply indicates a question; its absence, a statement.

4.3.7 Performance

The feature set of e-*motif* is summarised in Table 4.8. In order to test the accuracy of the model in predicting these features from text, 100 sentences are selected from the "oz06: 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 by the author 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.

Table 4.8: e-*motif* feature set

Feature	**Values**
Cognitive	
Judgment	000, 001, 010, 011, 100, 101, 110, 111
Focus	000, 001, 010, 011, 100, 101, 110, 111
Tense	past, present, future
Social	
Power	00, 01, 10, 11, narrative
Interaction	adaptation, coordination, regulation, narrative
Rhetoric	statement, question

The automatically predicted feature values are compared against the gold standard on the sentence level to produce the accuracies in Table 4.9 (one sentence corresponds to one percentage point). 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.

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 despite a few cases of ambiguity (future tense versus modal tense), 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% on general data (Section 2.2.2). Inspection of a subset of 33 discourse entity mentions across the 100 test sentences revealed a coreference resolution accuracy of 66% as an approximate for the children's literature domain. 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.

Time did not allow any optimisation of the feature assignment in this theory-driven research effort.

Table 4.9: e-*motif* feature 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		

Furthermore, a performance comparison to the model of Shaikh et al. (2009a) is not possible, since their model produces concrete surface emotion categories, whereas e-*motif* produces anonymous linguistic features for the purpose of implicit prosodic modelling.

4.4 Conclusion

e-*motif* is a model of affect that predicts antecedental features in the linguistic domain for use in the prosodic domain. These are the cognitive features of judgment, focus and tense, which are based on the OCC model, and the social features of power, interaction and rhetoric. It is important to note that, contrary to the OCC model, e-*motif* does not commit to discrete consequential emotions, but produces only anonymous affective states. This is necessary to support its subjective nature, that is its ability to model affect according to any subjective belief system (as long as the person remains consistent in it). This anonymity is not a limitation, because the model is intended as an implicit instrument for the prosodic domain to improve the naturalness of synthesised speech, in a way that measures the latter as closer to (the personality of, or speaker choice in) the original speech.

In order to be grounded in prosodic theory, particularly (cognitive and social) emotion theory, in a more principled way, e-*motif* employs the higher linguistic levels of discourse and information structure in its compound feature calculation. Character mentions in the audiobook narrative can be chained to assign valence in a consistent way for the features of judgment and power. A very important aspect of the OCC model, its focus of attention, can be mapped reliably with information structure to the feature of focus on the characters participating in the AGENT-verb-PATIENT semantics of an event. However, that which is the theoretical strength of e-*motif* also appears to be its practical weakness, since the state-of-the-art coreference resolution of Stanford CoreNLP is not accurate enough to produce usable strictly correct compound features. This disproves the first hypothesis of this thesis.

Chapter 5 will next investigate the acoustic effects of the e-*motif* features in audiobook speech and discuss the ramifications of the low predictive ability of the model regarding the research hypothesis that it should model prosodic behaviour in speech better.

Chapter 5

The Prosody of Affect

5.1 Introduction

A subjective model of affect, called e-*motif*, was proposed in the previous chapter to predict linguistic markers for prosodic behaviour in speech. This chapter will examine the proficiency of e-*motif* to account for prosody in natural speech based on the effects of the linguistic features on acoustic measures from the audiobook test set, and compare them to the statistics mentioned in Shaikh et al. (2010).

Shaikh et al. (2010) define empirically-motivated rules to map the consequential discrete emotions of their model to explicit acoustic parameters for their TTS system. e-*motif* takes the conservative approach of letting the antecedental linguistic features separate the acoustic data in the HTS framework appropriately, in order to model prosody implicitly. Crucially though, it allows speaker idiosyncrasies in the prosodic manifestation (albeit only if they are systematic). The chapter will discuss the success of this approach by comparing the synthesised speech of a baseline voice and the e-*motif* voice, both trained on the audiobook training set, to the original natural speech of the narrator in the audiobook test set.

5.2 Affective Prosody in Natural Speech

Shaikh et al. (2010) examine the changes in speech rate (*duration*), pitch average (*F0 average*), pitch range (*F0 minimum and maximum*) and *intensity* when they compare emotional natural speech to neutral natural speech. The effects of the six basic emotions of joy, sadness, fear, anger, disgust and surprise match those which are noted in previous literature (Murray and Arnott, 1993). 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 same aforementioned acoustic 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 (Boslaugh and Watters, 2008):

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{5.1}$$

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 (Table A.1 in Appendix A). The traditional significance level of $p < 0.05$ is adjusted for non-direction (two-tailedness) and Bonferroni-corrected for the 30 binary features to $p < \frac{0.05}{2 \times 30} \approx 0.001$. The degrees of freedom are calculated as:

$$df = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\left(\frac{\sigma_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\sigma_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{5.2}$$

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 (Quené, 2007; Kollmeier et al., 2008).

Since the automatically calculated features have a low accuracy (Section 4.3.7) that can influence the interpretation of the effects, the gold standard features of the 100 sentence test subset are also evaluated. The acoustics are measured on the phonetic level and only segments that fall under AGENT-verb-PATIENT semantics are considered. *This results in 128481 automatic and 1824 gold standard segments to be tested in the Phil Chenevert speech, and 132870 automatic and 1824 gold standard segments in the Judy Bieber speech* (the slight difference in number is due to the quality control filtering). The duration values of the segments are available from the alignment information; the F0 and intensity values are extracted with Praat (Boersma and Weenink).

The following subsections analyse the outcomes of the experimentation on the Phil Chenevert and Judy Bieber speech, using the automatic and gold standard linguistic features, by way of detailed tables. Each table lists the sample distribution means of a particular acoustic measure ("Duration", "Average F0" and "Average Intensity") for each binary linguistic feature when the latter is deactivated ("off") and activated ("on"). For F0 and intensity, minimum, maximum and average values are tested, but only the average values are shown, since they are sufficiently representative of (very similar to) those of the rest. The difference ("diff") between the means and its t-statistic with degrees of freedom ("t (df)") follow. The calculated degrees of freedom approximate infinity for most cases, so the threshold t-value is 3.090. 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. Values of "nan" (not a number) indicate that no segments were found for a particular activated feature.

5.2.1 Phil Chenevert Speech/Automatic Linguistic Features

Table 5.1 shows the effects of the automatically calculated linguistic features on the *duration* of the Phil Chenevert speech in the full test set. The expectation is to see all (or at least most of) the judgment, focus, tense and power features yield an effect, since they are dependent in the e-*motif* appraisal of an anonymous emotion. In reality, only power_{00} has an increasing influence that is both statistically significant and perceptually distinguishable. From the table it can be seen that the influences of some of the judgment, tense and other power features, and all of the interaction features, are statistically significant but not larger than or equal to the JND. None of the focus features show an effect on duration; neither do the very accurate rhetoric features, surprisingly, since phrase/clause finality usually has strong prosodic indicators (Section 2.4.1).

If only singular effects are prominent, they might be spurious, indicating the presence of confounding factors. In the case of power_{00}, it might be useful to note that the antagonists (with bad/0 valences) in the "oz06: The Emerald City of Oz" test set are predominantly male, whereas the protagonists (with good/1 valences) are mostly female. The speaker could have chosen to speak more slowly (increased duration) in an exaggerated portrayal of a conversation between two antagonistic males. It is necessary to compare the F0 and intensity results to make this conclusive.

The effects on *average F0* are much more pronounced than on duration. Of 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 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 $\text{interaction}_{narrative}$) than for direct speech dialogue of the story characters (represented by the other interaction features). Within direct speech, the tone rises substantially in initial segments after the indirect speech narrative ($\text{interaction}_{adaptation}$), but falls and settles in subsequent segments ($\text{interaction}_{coordination}$ and $\text{interaction}_{regulation}$). For F0, the rhetoric features now come into their right by correctly modelling statements with a downstep 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 $\text{tense}_{present}$ and tense_{future} in direct speech dialogue. The power features exhibit the same behaviour, since they all occur in direct speech dialogue, except for $\text{power}_{narrative}$. It is worth noting, though, that power_{00} has a much smaller increasing effect, compared to the others. A lower than usual tone could support the male conversation hypothesis above. It might now also be pertinent to ask the question of whether power_{10} and power_{11} show increases in tone because of the speaker choice regarding direct speech, or because the protagonists are mostly female, necessitating a higher tone. It is probably a bit of both factors, since the interaction features, which subsume both protagonistic and antagonistic direct speech, show consistent increases.

The *average intensity* values in the table reveal that most features do not have an influence, even though some show statistically significant differences. Of the cognitive features, only tense_{future} increases the intensity perceptually, though very subtly. It might indicate emphasis on the *will* to do something, rather than its temporal aspect. A direct speech dialogue turn (represented by $\text{interaction}_{adaptation}$ and $\text{interaction}_{coordination}$) also has a slight increasing effect. Finally, a lower intensity is used for statements than for questions.

5.2.2 Phil Chenevert Speech/Gold Standard Linguistic Features

As mentioned earlier, the automatic prediction of the judgment, focus and power feature values is not very accurate. Hence, it is imperative to test the effects of the gold standard feature values as well, in order to validate the interpretations of the features in the automatic case. However, there is a trade-off. Whereas the accuracy of the automatically calculated features is in question, the robustness that the large sample size of the full test set (128481 segments spread among 30 binary features) lends to the statistics, is not. Conversely, the accuracy of the gold standard features is not in question, but the sample size of the gold standard test subset (1824 segments spread among 30 binary features) might suffer from sparsity problems and not be representative enough. Nevertheless, agreement on substantial trends should hopefully be reached.

The effects on the *duration* of the Phil Chenevert speech in the gold standard test subset can be seen in Table 5.2. The generally insignificant behaviour concurs with that in the automatic assessment. The features that do stand out here are judgment_{001}, judgment_{010} and power_{01}. The opposing impacts of the

former two might be explained if a judicial viewpoint is once again assumed. They then convey two strong stereotypically antithetical affective states—a bad agent doing a bad deed to a good patient versus a bad agent doing a good deed to another bad patient—that produce contrasting surface emotions. The reason behind the singular effect of power_{01} is unclear, since, although it represents a male antagonist dialogue turn, it has the opposite effect than in the automatic case (for power_{00}).

The *average F0* values show again that most of the activity happens in this domain. The prominent effects of judgment_{000}, judgment_{001} and focus_{010} are disparate to those in the automatic case, with no immediate spurious explanations available. Nevertheless, those influences of the tense, power, interaction and rhetoric features that are statistically significant and perceptually distinguishable agree with their automatic counterparts.

The impact of the gold standard features on the *average intensity* measures is very little, as in the automatic case. The rhetoric features are consistent. The singular effect of focus_{111} might be due to its discourse-new function, as noted in the previous subsection. However, here it causes an increase in intensity, whereas it causes a decrease in F0.

5.2.3 Judy Bieber Speech/Automatic Linguistic Features

A different speaker can potentially manifest prosody in different ways. Therefore, the effects of the linguistic features are also tested on the Judy Bieber speech.

Table 5.3 indicates that the automatically calculated linguistic features have no discernible effect on the *duration* of the Judy Bieber speech in the full test set. Some of the durational differences are statistically significant, but smaller than the JND.

Average F0 tells a completely different story. There are prominent effects almost across the board. 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}, $\text{power}_{narrative}$ and $\text{interaction}_{narrative}$ decrease F0 as a result of indirect speech narrative, whereas $\text{interaction}_{adaptation}$ and $\text{interaction}_{coordination}$ increase F0 to denote direct speech dialogue turns. Again, power_{00} has a smaller increasing effect than power_{01} and power_{10}, possibly according to the male antagonistic conversation conjecture.

Why $\text{tense}_{present}$ does not match tense_{future}, and $\text{interaction}_{regulation}$ does not match $\text{interaction}_{adaptation}$ and $\text{interaction}_{coordination}$ in increased F0 to indicate direct speech dialogue, is uncertain. Perhaps the definition of $\text{interaction}_{regulation}$ lends a clue—continued direct speech impersonation might cause fatigue that forces the speaker to revert to her normal style of voice. The idiosyncratic null effect of power_{11} is also unclear. Finally, the rhetoric features are in accordance with their Phil Chenevert counterparts.

The *average intensity* values show the same results as those for duration—some statistically significant differences, but no perceptually discernible ones.

5.2.4 Judy Bieber Speech/Gold Standard Linguistic Features

The effects of the gold standard linguistic features on the Judy Bieber speech in the test subset are found to be generally insignificant, according to Table 5.4. For *duration* and *average intensity* it is unsurprising if compared to the automatic situation, but for *average F0* it is surprising.

The features of power_{00} and power_{11} do show a consistent influence in the speech across duration and F0, most likely because of the confounding gender factors. Male conversation is enacted with a slower tempo and lower tone, whereas female conversation is enacted with a faster tempo and a higher tone. The substantial effect here of power_{00} on F0 confirms the automatic case, but the effect of power_{11} is contrary to the null effect in the automatic full test set. The latter might now be explained by too many impure (inaccurate) samples of power_{11} in the full test set that smooth out the gender effect—power_{11} is indeed composed of

Table 5.1: *t*-tests on the means of the acoustic measures for the *automatic* linguistic features, from the *Phil Chenevert* speech of the *full* test set (128481 **A***v***P** segments)

Linguistic Feature	**Acoustic Measure Means**											
	Duration (ms)				**Average F0 (Hz)**				**Average Intensity (dB)**			
	off	**on**	**diff**	*t* (*df*)	**off**	**on**	**diff**	*t* (*df*)	**off**	**on**	**diff**	*t* (*df*)
judgment_{000}	95.153	96.683	1.529	1.409 (inf)	126.428	127.764	1.336	1.120 (inf)	65.757	66.008	0.251	1.659 (inf)
judgment_{001}	95.371	92.488	*-2.884*	*4.081 (inf)*	126.584	124.640	-1.944	2.515 (inf)	65.744	66.074	*0.330*	*3.268 (inf)*
judgment_{010}	95.150	95.750	0.600	0.899 (inf)	126.476	126.334	-0.142	0.198 (inf)	65.738	66.083	*0.345*	*3.765 (inf)*
judgment_{011}	95.649	92.731	*-2.918*	*6.204 (inf)*	127.081	123.121	**-3.960**	**7.369 (inf)**	65.772	65.719	-0.053	0.797 (inf)
judgment_{100}	95.193	95.281	0.088	0.069 (inf)	126.415	128.787	2.372	1.745 (inf)	65.766	65.671	-0.095	0.543 (inf)
judgment_{101}	95.227	94.953	-0.273	0.509 (inf)	126.331	127.504	1.173	1.937 (inf)	65.750	65.873	0.123	1.600 (inf)
judgment_{110}	95.001	96.660	1.659	2.979 (inf)	126.218	128.334	**2.116**	**3.558 (inf)**	65.735	65.986	*0.251*	*3.367 (inf)*
judgment_{111}	94.635	95.945	*1.310*	*3.691 (inf)*	126.080	126.979	0.899	2.289 (inf)	65.903	65.579	*-0.324*	*6.619 (inf)*
focus_{000}	95.193	96.096	0.903	0.294 (inf)	126.477	122.166	-4.311	1.087 (inf)	65.767	64.887	-0.880	1.834 (inf)
focus_{001}	95.251	94.012	-1.239	1.509 (inf)	126.502	125.697	-0.805	0.866 (inf)	65.781	65.408	*-0.373*	*3.255 (inf)*
focus_{010}	95.209	95.010	-0.198	0.285 (inf)	126.305	128.666	2.361	2.956 (inf)	65.760	65.824	0.064	0.669 (inf)
focus_{011}	94.944	96.418	1.474	3.038 (inf)	126.178	127.865	**1.687**	**3.280 (inf)**	65.754	65.816	0.062	0.955 (inf)
focus_{100}	95.187	96.667	1.480	0.669 (inf)	126.437	131.169	4.732	1.722 (inf)	65.766	65.515	-0.251	0.817 (inf)
focus_{101}	95.199	95.145	-0.054	0.075 (inf)	126.288	129.062	**2.774**	**3.504 (inf)**	65.772	65.641	-0.131	1.334 (inf)
focus_{110}	95.087	96.276	1.189	1.923 (inf)	126.423	126.888	0.465	0.686 (inf)	65.745	65.958	0.214	2.533 (inf)
focus_{111}	95.751	94.747	-1.004	2.840 (inf)	127.746	125.430	**-2.316**	**5.937 (inf)**	65.770	65.759	-0.010	0.215 (inf)
tense_{past}	96.815	93.988	*-2.827*	*7.935 (inf)*	131.802	122.486	**-9.316**	**23.710 (inf)**	66.078	65.530	*-0.547*	*11.214 (inf)*
$\text{tense}_{present}$	94.116	96.891	*2.775*	*7.656 (inf)*	123.462	131.183	**7.721**	**19.275 (inf)**	65.611	66.005	*0.394*	*7.977 (inf)*
tense_{future}	95.162	96.041	0.879	0.930 (inf)	126.003	138.117	**12.114**	**12.432 (inf)**	65.722	66.816	**1.093**	**8.742 (inf)**
power_{00}	94.922	100.155	**5.233**	**6.275 (inf)**	126.262	130.142	**3.879**	**4.388 (inf)**	65.738	66.232	*0.494*	*4.629 (inf)*
power_{01}	95.086	96.556	1.470	2.104 (inf)	125.621	136.993	**11.372**	**14.713 (inf)**	65.715	66.376	*0.661*	*7.162 (inf)*
power_{10}	95.015	97.245	*2.231*	*3.319 (inf)*	125.510	137.302	**11.792**	**16.777 (inf)**	65.704	66.443	*0.739*	*8.573 (inf)*
power_{11}	94.792	96.868	*2.076*	*4.564 (inf)*	124.122	136.193	**12.071**	**23.960 (inf)**	65.668	66.163	*0.495*	*8.063 (inf)*
$\text{power}_{narrative}$	97.314	93.774	*-3.541*	*9.768 (inf)*	135.777	120.218	**-15.559**	**39.169 (inf)**	66.268	65.426	*-0.842*	*17.102 (inf)*
$\text{interaction}_{adaptation}$	95.068	98.106	*3.039*	*3.353 (inf)*	125.687	144.235	**18.549**	**17.710 (inf)**	65.712	66.961	**1.249**	**10.744 (inf)**
$\text{interaction}_{coordination}$	94.988	99.499	*4.511*	*5.031 (inf)*	125.893	138.346	**12.454**	**13.285 (inf)**	65.711	66.862	**1.150**	**10.028 (inf)**
$\text{interaction}_{regulation}$	94.422	96.888	*2.466*	*6.390 (inf)*	122.900	134.269	**11.368**	**26.958 (inf)**	65.616	66.088	*0.472*	*9.068 (inf)*
$\text{interaction}_{narrative}$	97.314	93.774	*-3.541*	*9.768 (inf)*	135.777	120.218	**-15.559**	**39.169 (inf)**	66.268	65.426	*-0.842*	*17.102 (inf)*
$\text{rhetoric}_{statement}$	92.611	95.272	2.662	2.588 (inf)	141.659	126.011	**-15.648**	**13.561 (inf)**	66.976	65.728	**-1.249**	**9.110 (inf)**
$\text{rhetoric}_{question}$	95.272	92.611	-2.662	2.588 (inf)	126.011	141.659	**15.648**	**13.561 (inf)**	65.728	66.976	**1.249**	**9.110 (inf)**

Table 5.2: *t*-tests on the means of the acoustic measures for the *gold standard* linguistic features, from the *Phil Chenevert* speech of the test *sub*set (1824 **A***v***P** segments)

Linguistic Feature	**Acoustic Measure Means**											
	Duration (ms)				**Average F0 (Hz)**				**Average Intensity (dB)**			
	off	**on**	**diff**	*t* (*df*)	**off**	**on**	**diff**	*t* (*df*)	**off**	**on**	**diff**	*t* (*df*)
judgment_{000}	99.844	103.258	3.414	0.473 (098)	137.038	103.055	**-33.984**	**5.589 (inf)**	66.017	64.474	-1.542	1.492 (095)
judgment_{001}	101.304	81.967	**-19.337**	**4.374 (inf)**	136.728	116.577	**-20.151**	**3.420 (inf)**	66.076	64.062	-2.014	2.597 (inf)
judgment_{010}	97.884	113.098	**15.214**	**3.094 (inf)**	133.355	147.841	14.486	2.501 (inf)	65.808	66.760	0.951	1.628 (inf)
judgment_{011}	99.291	102.701	3.410	0.804 (inf)	135.880	133.510	-2.370	0.516 (inf)	65.951	65.904	-0.047	0.092 (inf)
judgment_{100}	100.539	90.619	-9.920	1.509 (inf)	136.297	119.059	-17.238	1.919 (inf)	66.044	64.111	-1.934	2.031 (inf)
judgment_{101}	100.378	94.057	-6.322	1.307 (inf)	134.859	143.821	8.962	1.090 (inf)	65.930	66.132	0.202	0.221 (inf)
judgment_{110}	99.994	100.171	0.177	0.030 (inf)	134.800	140.846	6.046	0.847 (inf)	66.042	64.995	-1.047	1.367 (inf)
judgment_{111}	100.781	98.420	-2.361	0.719 (inf)	133.390	139.490	6.099	1.561 (inf)	65.544	66.763	1.219	2.982 (inf)
focus_{000}	100.011	nan	nan	nan (nan)	135.380	nan	nan	nan (nan)	65.941	nan	nan	nan (nan)
focus_{001}	99.582	103.094	3.513	0.669 (inf)	136.830	124.973	-11.856	2.136 (inf)	66.154	64.415	-1.739	2.711 (inf)
focus_{010}	100.134	99.579	-0.555	0.147 (inf)	131.948	147.444	**15.496**	**3.522 (inf)**	65.889	66.126	0.237	0.467 (inf)
focus_{011}	99.852	100.243	0.390	0.121 (inf)	135.667	134.962	-0.705	0.180 (inf)	65.976	65.890	-0.086	0.208 (inf)
focus_{100}	100.028	99.302	-0.726	0.074 (044)	134.720	162.719	27.999	2.326 (044)	65.933	66.288	0.355	0.251 (044)
focus_{101}	100.139	89.048	-11.091	1.324 (022)	135.413	132.553	-2.860	0.184 (021)	65.963	64.100	-1.863	1.013 (021)
focus_{110}	100.389	91.200	-9.189	1.341 (083)	135.886	123.576	-12.310	1.553 (084)	65.994	64.720	-1.273	1.229 (080)
focus_{111}	99.854	100.759	0.905	0.216 (inf)	137.020	127.553	-9.467	2.095 (inf)	65.664	67.267	**1.604**	**3.340 (inf)**
tense_{past}	97.915	104.032	6.117	1.819 (inf)	140.434	125.685	**-14.750**	**3.904 (inf)**	66.246	65.358	-0.888	2.065 (inf)
$\text{tense}_{present}$	99.396	100.929	1.533	0.475 (inf)	131.052	141.836	10.784	2.723 (inf)	65.987	65.874	-0.113	0.273 (inf)
tense_{future}	102.358	93.191	-9.168	2.578 (inf)	134.397	138.236	3.839	0.948 (inf)	65.636	66.829	1.193	2.617 (inf)
power_{00}	97.636	107.662	10.026	2.434 (inf)	136.933	130.375	-6.558	1.516 (inf)	65.795	66.414	0.620	1.231 (inf)
power_{01}	100.355	69.000	**-31.355**	**5.330 (022)**	135.534	121.485	-14.049	0.727 (019)	65.940	66.084	0.145	0.075 (019)
power_{10}	99.983	102.500	2.517	0.208 (020)	135.277	144.709	9.432	0.388 (019)	65.948	65.317	-0.632	0.337 (019)
power_{11}	103.911	95.809	-8.103	2.574 (inf)	124.165	147.464	**23.300**	**6.248 (inf)**	65.861	66.028	0.167	0.412 (inf)
$\text{power}_{narrative}$	99.304	102.025	2.722	0.761 (inf)	141.570	117.751	**-23.819**	**6.239 (inf)**	66.142	65.370	-0.772	1.679 (inf)
$\text{interaction}_{adaptation}$	100.960	95.431	-5.528	1.448 (inf)	130.786	157.558	**26.772**	**4.930 (inf)**	65.721	67.005	1.284	2.559 (inf)
$\text{interaction}_{coordination}$	101.713	97.066	-4.647	1.420 (inf)	133.526	138.589	5.063	1.264 (inf)	65.839	66.119	0.280	0.653 (inf)
$\text{interaction}_{regulation}$	98.330	106.640	8.310	1.942 (inf)	135.881	133.406	-2.475	0.578 (inf)	66.065	65.452	-0.613	1.204 (inf)
$\text{interaction}_{narrative}$	99.304	102.025	2.722	0.761 (inf)	141.570	117.751	**-23.819**	**6.239 (inf)**	66.142	65.370	-0.772	1.679 (inf)
$\text{rhetoric}_{statement}$	92.541	100.854	8.314	1.832 (inf)	161.896	132.387	**-29.509**	**5.137 (inf)**	68.471	65.656	**-2.815**	**4.724 (inf)**
$\text{rhetoric}_{question}$	100.854	92.541	-8.314	1.832 (inf)	132.387	161.896	**29.509**	**5.137 (inf)**	65.656	68.471	**2.815**	**4.724 (inf)**

good/1 valences, which are also the default values for characters when the coreference resolution fails. This then also accounts for the more pronounced power_{11} effect on F0 in the Phil Chenevert speech when the gold standard values are used instead of the automatically calculated ones (Section 5.2.2 versus 5.2.1).

The only other gold standard feature that has a noticeable impact in the Judy Bieber speech is focus_{111} with an increase in F0, as opposed to the decrease it causes in the Phil Chenevert speech.

5.2.5 Summary

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. Furthermore, gender impersonation has an overriding effect on the power features as well. Finally, the inaccuracy of power_{11} in particular is not completely detrimental, but only lessens its effect on the speech.

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

5.3 Affective Prosody in Synthesised Speech

Shaikh et al. (2010) 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 5.5 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. However, the positional features take on a relative, ordinal-type format, instead of the usual cardinal-type one. This "initial, medial and final" segmentation reduces the number of overall features to be modelled, yet is still prosodically robust (Féry, 2009, 2010; Tseng, 2010). 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).

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.

Table 5.3: t-tests on the means of the acoustic measures for the *automatic* linguistic features, from the *Judy Bieber* speech of the *full* test set (132870 **A**v**P** segments)

Linguistic Feature	**Acoustic Measure Means**											
	Duration (ms)				**Average F0 (Hz)**				**Average Intensity (dB)**			
	off	**on**	**diff**	t (df)	**off**	**on**	**diff**	t (df)	**off**	**on**	**diff**	t (df)
judgment_{000}	82.499	81.458	-1.040	1.209 (inf)	236.175	244.876	**8.701**	**5.182 (inf)**	65.533	65.613	0.080	0.484 (inf)
judgment_{001}	82.538	81.398	-1.140	2.033 (inf)	236.247	238.888	2.641	2.368 (inf)	65.537	65.506	-0.031	0.273 (inf)
judgment_{010}	82.526	81.786	-0.740	1.467 (inf)	236.119	239.938	**3.819**	**3.737 (inf)**	65.534	65.543	0.008	0.082 (inf)
judgment_{011}	82.526	82.178	-0.348	0.904 (inf)	237.085	232.710	**-4.375**	**5.924 (inf)**	65.624	65.053	*-0.571*	*7.650 (inf)*
judgment_{100}	82.515	80.408	-2.106	2.137 (inf)	236.211	245.506	**9.295**	**4.935 (inf)**	65.534	65.587	0.053	0.282 (inf)
judgment_{101}	82.450	82.632	0.182	0.419 (inf)	236.180	238.119	1.939	2.338 (inf)	65.526	65.606	0.080	0.955 (inf)
judgment_{110}	82.511	82.177	-0.334	0.788 (inf)	235.979	239.551	**3.571**	**4.372 (inf)**	65.513	65.701	0.189	2.334 (inf)
judgment_{111}	82.041	83.040	*0.999*	*3.566 (inf)*	237.830	234.518	**-3.312**	**6.210 (inf)**	65.455	65.640	*0.185*	*3.436 (inf)*
focus_{000}	82.455	88.095	5.640	1.989 (inf)	236.467	213.834	**-22.633**	**4.445 (inf)**	65.539	64.169	-1.370	2.863 (inf)
focus_{001}	82.352	84.997	*2.645*	*3.794 (inf)*	237.056	222.584	**-14.472**	**11.448 (inf)**	65.579	64.602	*-0.977*	*8.101 (inf)*
focus_{010}	82.428	83.073	0.645	1.150 (inf)	236.523	234.711	-1.812	1.749 (inf)	65.537	65.511	-0.026	0.246 (inf)
focus_{011}	82.615	81.766	-0.849	2.289 (inf)	236.056	238.096	2.039	2.924 (inf)	65.489	65.759	*0.269*	*3.844 (inf)*
focus_{100}	82.447	86.375	3.928	1.932 (inf)	236.457	227.606	-8.851	2.559 (inf)	65.535	65.485	-0.050	0.153 (inf)
focus_{101}	82.332	84.555	*2.223*	*3.667 (inf)*	236.893	229.083	**-7.810**	**7.418 (inf)**	65.555	65.229	*-0.326*	*3.097 (inf)*
focus_{110}	82.351	83.698	1.347	2.667 (inf)	236.186	238.609	2.422	2.610 (inf)	65.512	65.769	0.257	2.765 (inf)
focus_{111}	83.183	81.907	*-1.277*	*4.553 (inf)*	234.525	237.892	**3.368**	**6.338 (inf)**	65.516	65.550	0.034	0.631 (inf)
tense_{past}	83.116	81.976	*-1.140*	*4.078 (inf)*	237.690	235.414	**-2.275**	**4.264 (inf)**	65.650	65.447	*-0.203*	*3.801 (inf)*
$\text{tense}_{present}$	81.789	83.509	*1.720*	*6.056 (inf)*	236.218	236.684	0.466	0.861 (inf)	65.497	65.593	0.096	1.784 (inf)
tense_{future}	82.608	78.992	*-3.616*	*5.177 (inf)*	235.937	248.252	**12.314**	**9.419 (inf)**	65.507	66.245	*0.738*	*5.597 (inf)*
power_{00}	82.459	82.686	0.227	0.379 (inf)	236.112	241.594	**5.482**	**4.624 (inf)**	65.505	66.075	*0.570*	*4.940 (inf)*
power_{01}	82.468	82.512	0.044	0.083 (inf)	235.759	244.220	**8.461**	**8.373 (inf)**	65.523	65.677	0.153	1.538 (inf)
power_{10}	82.449	82.725	0.276	0.534 (inf)	235.593	245.441	**9.848**	**10.074 (inf)**	65.526	65.633	0.107	1.119 (inf)
power_{11}	82.667	81.691	-0.977	2.892 (inf)	236.165	237.347	1.182	1.784 (inf)	65.480	65.754	*0.275*	*4.170 (inf)*
$\text{power}_{narrative}$	82.177	82.678	0.501	1.789 (inf)	240.777	233.338	**-7.439**	**13.805 (inf)**	65.757	65.379	*-0.378*	*7.013 (inf)*
$\text{interaction}_{adaptation}$	82.573	80.313	*-2.261*	*3.864 (inf)*	235.483	255.809	**20.326**	**14.954 (inf)**	65.502	66.237	*0.735*	*5.549 (inf)*
$\text{interaction}_{coordination}$	82.625	79.463	*-3.161*	*5.734 (inf)*	235.536	253.391	**17.855**	**13.708 (inf)**	65.502	66.172	*0.669*	*5.315 (inf)*
$\text{interaction}_{regulation}$	82.292	82.855	0.563	1.873 (inf)	236.256	236.717	0.462	0.818 (inf)	65.493	65.626	0.133	2.363 (inf)
$\text{interaction}_{narrative}$	82.177	82.678	0.501	1.789 (inf)	240.777	233.338	**-7.439**	**13.805 (inf)**	65.757	65.379	*-0.378*	*7.013 (inf)*
$\text{rhetoric}_{statement}$	79.238	82.583	*3.345*	*4.803 (inf)*	243.336	236.163	**-7.173**	**4.961 (inf)**	66.047	65.517	*-0.530*	*3.534 (inf)*
$\text{rhetoric}_{question}$	82.583	79.238	*-3.345*	*4.803 (inf)*	236.163	243.336	**7.173**	**4.961 (inf)**	65.517	66.047	*0.530*	*3.534 (inf)*

Table 5.4: *t*-tests on the means of the acoustic measures for the *gold standard* linguistic features, from the *Judy Bieber* speech of the test *sub*set (1824 **A***v***P** segments)

Linguistic Feature	**Acoustic Measure Means**											
	Duration (ms)				**Average F0 (Hz)**				**Average Intensity (dB)**			
	off	**on**	**diff**	*t* (*df*)	**off**	**on**	**diff**	*t* (*df*)	**off**	**on**	**diff**	*t* (*df*)
judgment_{000}	83.510	82.360	-1.151	0.229 (098)	239.732	227.629	-12.104	1.137 (098)	66.081	66.748	0.667	0.658 (099)
judgment_{001}	83.784	78.852	-4.931	1.368 (inf)	239.525	233.801	-5.724	0.681 (inf)	66.112	66.131	0.019	0.022 (inf)
judgment_{010}	81.969	92.588	10.619	3.002 (inf)	241.775	222.939	-18.837	2.884 (inf)	66.059	66.445	0.385	0.613 (inf)
judgment_{011}	81.508	90.727	9.219	2.919 (inf)	242.228	227.606	-14.622	2.593 (inf)	66.079	66.241	0.162	0.276 (inf)
judgment_{100}	83.827	76.804	-7.023	1.714 (inf)	238.598	248.820	10.222	0.796 (inf)	66.234	63.969	-2.265	1.857 (inf)
judgment_{101}	83.574	81.509	-2.064	0.455 (inf)	237.462	266.367	28.905	2.300 (inf)	66.190	64.863	-1.327	1.400 (inf)
judgment_{110}	83.984	78.457	-5.527	1.511 (inf)	238.754	242.792	4.038	0.497 (inf)	66.122	66.028	-0.094	0.109 (inf)
judgment_{111}	85.688	78.840	-6.847	3.001 (inf)	234.434	248.866	14.432	2.855 (inf)	65.981	66.387	0.407	0.821 (inf)
focus_{000}	83.454	nan	nan	nan (nan)	239.142	nan	nan	nan (nan)	66.113	nan	nan	nan (nan)
focus_{001}	83.785	81.076	-2.709	0.672 (inf)	240.382	230.240	-10.142	1.332 (inf)	66.149	65.857	-0.291	0.427 (inf)
focus_{010}	83.732	82.475	-1.257	0.497 (inf)	236.506	248.405	11.899	2.045 (inf)	66.155	65.968	-0.187	0.340 (inf)
focus_{011}	81.932	85.674	3.742	1.619 (inf)	244.192	231.777	-12.414	2.545 (inf)	66.286	65.862	-0.424	0.871 (inf)
focus_{100}	83.436	84.186	0.750	0.106 (044)	239.577	221.116	-18.461	1.025 (044)	66.126	65.600	-0.525	0.359 (044)
focus_{101}	83.461	82.857	-0.604	0.082 (021)	238.813	267.343	28.530	1.161 (020)	66.140	63.839	-2.301	1.076 (020)
focus_{110}	83.676	78.267	-5.410	0.967 (081)	240.319	211.697	-28.621	2.278 (080)	66.152	65.219	-0.933	0.752 (080)
focus_{111}	83.687	82.342	-1.345	0.468 (inf)	235.197	257.965	**22.768**	**4.152 (inf)**	65.822	67.503	1.681	2.966 (inf)
tense_{past}	83.853	82.688	-1.165	0.492 (inf)	240.329	236.865	-3.464	0.717 (inf)	65.836	66.646	0.810	1.700 (inf)
$\text{tense}_{present}$	82.683	84.604	1.921	0.829 (inf)	239.926	237.972	-1.955	0.393 (inf)	66.413	65.666	-0.747	1.529 (inf)
tense_{future}	83.721	82.677	-1.045	0.407 (inf)	237.462	244.023	6.562	1.188 (inf)	66.117	66.102	-0.015	0.029 (inf)
power_{00}	80.682	92.384	**11.702**	**4.039 (inf)**	246.279	216.144	**-30.135**	**5.659 (inf)**	66.036	66.363	0.327	0.621 (inf)
power_{01}	83.331	94.500	11.169	1.270 (020)	238.783	271.524	32.741	0.906 (019)	66.128	64.778	-1.350	0.496 (019)
power_{10}	83.437	85.000	1.563	0.194 (020)	239.254	229.012	-10.242	0.402 (019)	66.140	63.717	-2.423	1.591 (020)
power_{11}	87.622	78.964	**-8.658**	**3.852 (inf)**	230.024	248.966	**18.942**	**3.954 (inf)**	66.592	65.598	-0.994	2.118 (inf)
$\text{power}_{narrative}$	83.578	83.101	-0.477	0.196 (inf)	238.502	240.965	2.463	0.496 (inf)	65.803	66.998	1.195	2.316 (inf)
$\text{interaction}_{adaptation}$	84.752	77.188	-7.563	2.929 (inf)	236.927	249.834	12.907	1.918 (inf)	66.204	65.676	-0.528	0.781 (inf)
$\text{interaction}_{coordination}$	83.270	83.772	0.503	0.217 (inf)	243.475	231.644	-11.831	2.294 (inf)	66.557	65.345	-1.212	2.477 (inf)
$\text{interaction}_{regulation}$	82.137	88.645	6.508	1.943 (inf)	238.593	241.304	2.711	0.486 (inf)	65.955	66.739	0.785	1.448 (inf)
$\text{interaction}_{narrative}$	83.578	83.101	-0.477	0.196 (inf)	238.502	240.965	2.463	0.496 (inf)	65.803	66.998	1.195	2.316 (inf)
$\text{rhetoric}_{statement}$	82.973	83.508	0.535	0.149 (inf)	234.275	239.691	5.416	0.770 (inf)	66.964	66.017	-0.947	1.191 (inf)
$\text{rhetoric}_{question}$	83.508	82.973	-0.535	0.149 (inf)	239.691	234.275	-5.416	0.770 (inf)	66.017	66.964	0.947	1.191 (inf)

Table 5.5: Features used in the HTS context labels

Feat	Description	Values
P context: syllable-level phonetic-verbal/suprasegmental features		
p_1	left triphone context (previous phone)	
p_2	centre 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 $(1, \leq 2, \leq 4, > 4)$
A context: word-level lexical-suprasegmental features		
a_1	syllable position in word	initial, medial, final
a_2	syllable count in word	isolated, short, medium, long $(1, \leq 2, \leq 4, > 4)$
a_3	syllable lexical function in word (lexical stress)	primary, secondary, none
B context: phrase-level syntactic-suprasegmental features		
b_1	syllable position in phrase	initial, medial, final
b_2	syllable count in phrase	isolated, short, medium, long $(1, \leq 4, \leq 8, > 8)$
b_3	word position in phrase	initial, medial, final
b_4	word count in phrase	isolated, short, medium, long $(1, \leq 2, \leq 4, > 4)$
b_5	word syntactic function in phrase (phrase stress)	head, modifier
C context: clause-level semantic-suprasegmental features		
c_1	syllable position in clause	initial, medial, final
c_2	syllable count in clause	isolated, short, medium, long $(1, \leq 12, \leq 24, > 24)$
c_3	phrase position in clause	initial, medial, final
c_4	phrase count in clause	isolated, short, medium, long $(1, \leq 2, \leq 3, > 3)$
c_5	phrase semantic function in clause	agent, verb, patient, other
D context: discourse-level cognitive/pragmatic-augmentative and affective 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-augmentative and affective 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

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—128481 Phil Chenevert speech segments and 132870 Judy Bieber speech segments. 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 (Boslaugh and Watters, 2008):

$$\chi^2 = \frac{(|n_1 - n_2| - 0.5)^2}{n_1 + n_2} \tag{5.3}$$

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 (Table A.2

in Appendix A). 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 following subsections discuss the results of the synthetic voice comparisons listed in tables and illustrated with figures. 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.

Each figure shows the sample allocations as portions in a pie chart. The reference synthetic voice portion ("Baseline" in the first case and "Cognitive" in the second) is always drawn in blue. The portion of the synthetic voice under evaluation ("Cognitive" in the first case and "Social" in the second) is drawn in green if it is significantly closer to the natural voice than the reference voice; red otherwise. The "Equal" portion is drawn in yellow.

5.3.1 Phil Chenevert Speech

Figure 5.1 (a), (b) and (c) illustrate the results of the comparison between the baseline and cognitive synthetic voices, for the Phil Chenevert speech. Exact numbers are available in the upper half of Table 5.6. The cognitive version shows no improvement over the baseline version for duration, F0 and intensity. This is not surprising, since the e-*motif* judgment and focus features generally have no discernible effect on the Phil Chenevert natural speech, and the tense feature effects are confounded by direct speech dialogue factors.

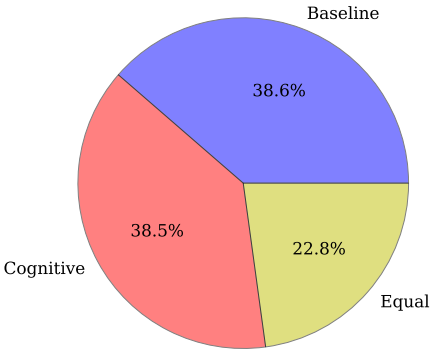
The cognitive-social comparison is portrayed in Figure 5.1 (d), (e) and (f), and the lower half of Table 5.6. The more robust social version, which models the strong differences between indirect speech narrative and direct speech dialogue, do improve F0 and intensity, which were shown to be the areas of activity of the effects (Section 5.2.1 and 5.2.2).

Table 5.6: McNemar comparisons between the synthetic voices on the full test set, for *Phil Chenevert*

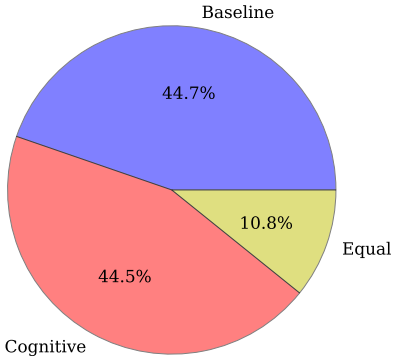
Measure	**A*v*P Segments**				χ^2
	Total	**Baseline**	**Cognitive**	**Equal**	
Duration	128481	49632	49510	29339	0.149
F0	128481	57450	57164	13867	0.711
Intensity	128481	63921	64559	1	3.163
Measure	**A*v*P Segments**				χ^2
	Total	**Cognitive**	**Social**	**Equal**	
Duration	128481	48291	47421	32769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

5.3.2 Judy Bieber Speech

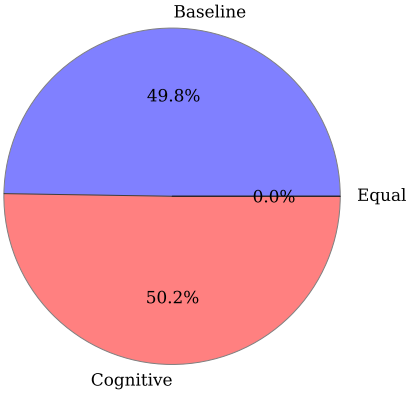
The results for the Judy Bieber synthetic voices can be viewed in Figure 5.2 and Table 5.7 There are no improvements when first the cognitive context (subfigures (a), (b) and (c) and the upper half of the table) and then the social context (subfigures (d), (e) and (f) and the lower half of the table) are added. This is as expected for duration and intensity, but not for F0, since the Judy Bieber natural speech is amenable to both the cognitive and the social features of e-*motif* in the F0 domain. 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



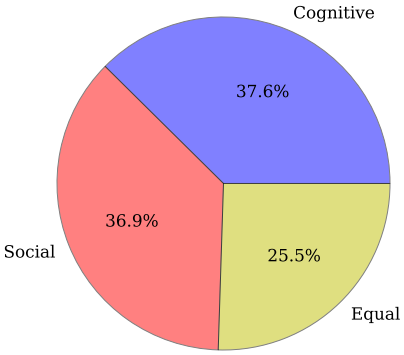
(a) Baseline-Cognitive Duration



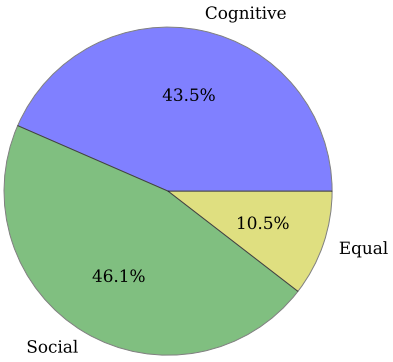
(b) Baseline-Cognitive F0



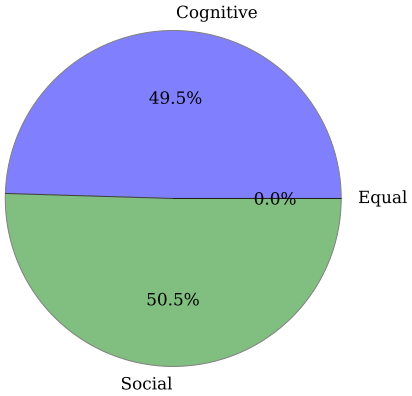
(c) Baseline-Cognitive Intensity



(d) Cognitive-Social Duration



(e) Cognitive-Social F0



(f) Cognitive-Social Intensity

Figure 5.1: McNemar comparisons between the synthetic voices on the full test set, for *Phil Chenevert*

e-*motif* features during the decision tree clustering process. The strength of these positional and counting features has been noted in a previous study (Schlünz et al., 2010).

Table 5.7: McNemar comparisons between the synthetic voices on the full test set, for *Judy Bieber*

Measure	**A*v*P 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	**A*v*P 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

5.3.3 Summary

The synthesised voices trained on the Phil Chenevert speech perform as expected, given the outcomes of the effects of the cognitive and social contexts on natural speech. The cognitive features do not contribute significantly enough to the quality of the HTS data separation process, but the social features do.

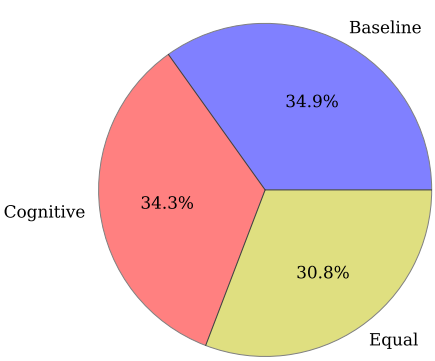
The Judy Bieber case is different for the worse, since the cognitive version of the synthetic voices is not an improvement over the baseline version, even though the cognitive features have an effect on the natural speech. Furthermore, the social version shows the same quality, despite the social features also being prominent in the natural speech. This is most likely due to the positional and counting features that model the Judy Bieber prosody with sufficient influence to override the other features.

5.4 Conclusion

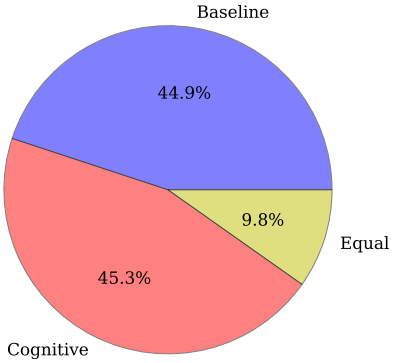
The experimental investigation in this chapter reveals a few important antitheses in the ability of e-*motif* to model prosodic behaviour in speech. e-*motif* is able to model the prosodic differences between indirect speech narrative and direct speech dialogue via the spurious effects of its social 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.

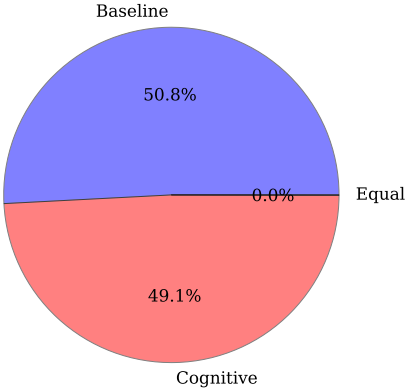
If the latter is true, the implication is then that prosodic phenomena can and need only be robustly explained by *superficial structure* at the current grain of NLP analysis—that is sentence-internal syntactic-like structure, and sentence-external dialogue structure. This will largely disprove the second research hypothesis. The next chapter will summarise the thesis work and make some closing remarks.



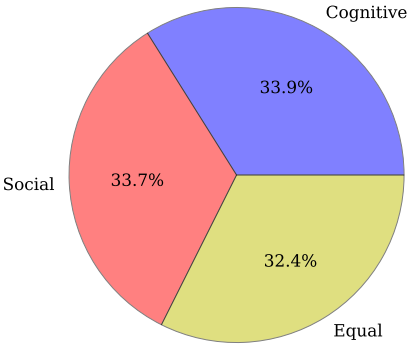
(a) Baseline-Cognitive Duration



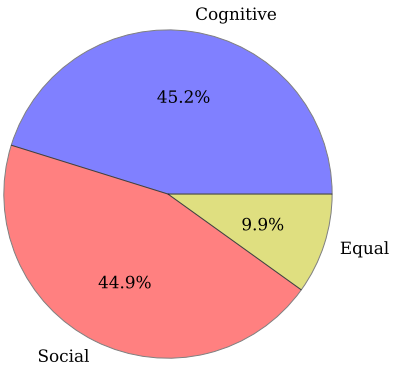
(b) Baseline-Cognitive F0



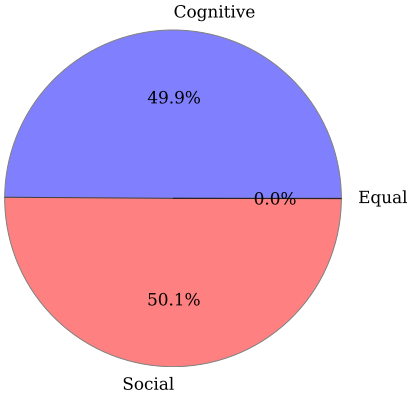
(c) Baseline-Cognitive Intensity



(d) Cognitive-Social Duration



(e) Cognitive-Social F0



(f) Cognitive-Social Intensity

Figure 5.2: McNemar comparisons between the synthetic voices on the full test set, for *Judy Bieber*

Chapter 6

Conclusion

6.1 In Retrospect

Speech synthesis plays a pivotal role in augmentative and alternative communication, being able to vocalise dynamic content. This enables the AAC user to construct any message for any situation. However, state-of-the-art synthesis systems still struggle on the front of naturalness. Human-like synthetic speech would be a preferable quality for AAC, since it should ideally embody that which its impeded biological counterpart cannot bring to the conversation.

Towards this end, the research set out to understand the factors of meaning and intent in communication that give rise to emphasis and emotion. In particular, two questions were asked in Chapter 1. The first question was on whether meaning, intent, emphasis and emotion can be predicted accurately enough from text using current text analysis tools, since speech synthesis must infer all communicative phenomena from plain text. The second question was on whether systematic acoustic correlates of these textual antecedents can be determined from human speech so that the naturalness of synthesised speech can be improved. It was hypothesised that both these questions could be answered in the affirmative, given the trends in literature.

The literature study in Chapter 2 formalised the disciplines of text analysis into NLP and speech synthesis into TTS. Important linguistic concepts were discussed with reference to their effect on the non-verbal component of speech, called prosody. It was noted that, in order to model prosody more appropriately, it is necessary to include the linguistic antecedents of discourse, information structure and affect in the text analysis. Discourse and information structure characterise meaning, intent and emphasis in the forms of themes and rhemes, givenness and focus. Affect formalises the modelling of emotion from text to speech. State-of-the-art NLP software can track discourse and information structure with shallow semantic parsing and coreference resolution, though do not yet employ these higher linguistic levels towards affect detection from text.

To address this shortcoming, Chapter 3 promoted the use of audiobook text and speech as a resource of discourse, information structure and affect because the narrative is more suitable to evoke these phenomena than the traditional TTS corpus. The Oz series of audiobooks by the writer L. Frank Baum and the speakers Phil Chenevert and Judy Bieber were chosen, since these children's stories have grammatical and moral simplifications that should aid in the NLP development and performance of a model of affect. The text and speech of the audiobooks were phonetically aligned, employing the quality control technique of DP scoring. Even though the DP scoring turned out not to be that well-suited to prosodically varying audiobook speech under the HTK framework, the alignments were found to be robust nonetheless.

Towards the first aim of developing an NLP system that can model discourse, information structure and affect, the work on affect by Shaikh et al. (2009a) was extended to the model e-*motif* in Chapter 4. Their initial implementation of the OCC model sought to take a step back from the surface level of emotional

expressions and rather identify the underlying factors that contribute towards them. e-*motif* builds on this by specifying the cognitive features of judgment, focus and tense, which are based on the OCC model, and the social features of power, interaction and rhetoric. The subjective nature of e-*motif* is an improvement on their work, in that the e-*motif* features do not commit to discrete consequential emotions, but produce anonymous affective states that are useful for prosodic modelling. e-*motif* is grounded further in prosodic theory by implementing the higher linguistic levels of discourse and information structure with the coreference resolution of Stanford CoreNLP. However, the theoretical strength of e-*motif* is also its pragmatic downfall, since the performance evaluation revealed that Stanford CoreNLP is not accurate enough to produce usable feature values. The first hypothesis of this thesis is thus disproved.

The second aim of the research was to find prosodic correlates of the e-*motif* antecedental features in the aligned natural audiobook speech, in order to improve synthesised speech. Chapter 5 showed that e-*motif* was able to distinguish between the prosody of indirect speech narrative and direct speech dialogue via the effects of its social features. It could be seen that Phil Chenevert preferred to use such prosody, with significant effects in his natural speech and significant impact on the data separation process in the HTS framework. On the contrary, Judy Bieber applied her tone in such a way that the naturally significant social features did not influence the quality of her synthesised speech. The cognitive features of e-*motif* showed effects in the evenly toned natural speech of Judy Bieber, but not the variably toned natural speech of Phil Chenevert. However, for Judy Bieber, the computationally much simpler positional and counting features in the HTS context labels can account for such prosody in her synthesised speech with similar quality as the complex e-*motif* features do.

The positional and counting information can be seen as a naive kind of syntactic structure. It is, therefore, prudent to ask whether the cognitive features do not rather show an effect in the natural speech because of confounding structural factors. If so, it may be concluded that, with the current state of NLP, prosodic phenomena can and need only be explained by superficial sentence-internal syntactic structure and sentence-external dialogue structure. This would disprove the second hypothesis and comment on a statement by Ferdinand de Saussure, one of the fathers of modern linguistics (de Saussure, 1916; Taylor, 2009)—we cannot yet make sense of "the arbitrary nature of form and meaning in communication".

6.2 In Prospect

The thesis touched on three broad areas of research upon which improvement might be shown in future work. These are the compilation and phonetic alignment of a corpus of natural language text and speech, the NLP analysis of the text to infer features for a model of affect, and the prosodic analysis of the model effects in natural and synthesised speech.

The narrative-structured audiobook is a very useful step up from the traditional randomly-constituted TTS corpus in capturing discourse-level linguistic and prosodic phenomena. This study focussed on the very narrow genre of children's literature for simplifying reasons. Nevertheless, it would be useful in future to evaluate e-*motif* across other genres as well to obtain wider prosodic coverage. Compared to such a more inclusive set of genres, the current performance measures are likely to be pessimistic, since it would be a relatively easier task to detect and adapt to differences in genres[1]. From a puristic AAC viewpoint, however, it can be argued that the speech planning by the narrator of well-formed sentences and paragraphs in well-structured monologues and dialogues is still somewhat artificial compared to true spontaneous conversational speech. A corpus of the latter should be beneficial towards the creation of more natural TTS voices for AAC. The phonetic alignment process is already robust, but it might be interesting to investigate the effects of using target data-specific acoustic models, even more advanced HMMs or DNNs.

[1] Thanks to the various examiners for their additional suggestions for future work.

On the NLP front, the semantic analysis of the text could be improved beyond the agent-verb-patient granularity by incorporating semantic role labelling. This technique takes the verb predicate and identifies not only the semantic agent and patient, but also other roles such as time, duration, place, manner, means and purpose. These finer-grained roles would allow the formulation of a more representative or comprehensive anonymous affective state. Of course, increasing the number of identifiable semantic roles decreases the attainable accuracy. The semantic role labelling would need to perform well enough, as is required from the coreference resolution, for a model like e-*motif* to work.

The investigation of natural speech prosody might deliver more systematic and insightful results if the effects of the cognitive and social features could somehow be analysed separately from those of the syntactic-structural features. The same goes for the separation of monologue and dialogue speech. In particular, the analysis could also benefit from both gold standard features and automatically extracted features being evaluated on the gold standard test subset. The cognitive and social synthetic voice versions would also have to be compared separately to the baseline version. An inspection of the HTS decision tree during voice building, in order to see how the structural features or the e-*motif* features are favoured, should be most informative. The quality of the synthesised speech might also be improved by using a TTS framework other than HTS, with a data separation technique other than a decision tree, in order to model the dependencies among the e-*motif* features in a more accountable way[1].

In the meantime, state-of-the-art TTS voices will have to suffice for AAC systems. In still not being able to account for all meaning and intent in speech and language, they cannot yet *imitate* human communication, but they can certainly *facilitate* it. Many a human listener will be able to "read between the lines" of an AAC user's synthesised speech, making the conversation edifying for both parties.

Bibliography

C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet project. In *Proceedings of COLING-ACL'98*, pages 86–90, Montreal, Canada, 1998.

C. Bartneck. Integrating the OCC model of emotions in embodied characters, 2002.

M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel. *Prosodic Typology: The Phonology of Intonation and Phrasing*, chapter The Original ToBI System and the Evolution of the ToBI Framework. Oxford University Press, 2006.

A. W. Black, H. Zen, and K. Tokuda. Statistical parametric speech synthesis. In *Proceedings of ICASSP 2007*, 2007.

P. Boersma and D. Weenink. Praat: doing phonetics by computer. `http://www.praat.org/`.

S. Boslaugh and P. A. Watters. *Statistics in a nutshell*. O'Reilly Media, Inc., first edition, 2008.

R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions On Affective Computing*, 1(1):18–37, 2010.

Y. Choi and C. Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801, 2008.

N. Chomsky. Minimalist inquiries: The framework. In R. Martin, D. Michaels, and J. Uriagereka, editors, *Step by Step: Essays in Minimalist Syntax in Honor of Howard Lasnik*, pages 89–115. Cambridge, MA: MIT Press, 2000.

N. Chomsky. Derivation by phase. In M. Kenstowicz, editor, *Ken Hale: A Life in Language*, pages 1–52. Cambridge, MA: The MIT Press, 2001.

N. Chomsky and M. Halle. *The Sound Pattern of English*. New York: Harper and Row, 1968.

M. Davel and E. Barnard. Pronunciation prediction with Default&Refine. *Computer Speech and Language*, 22(4):374–393, 2008.

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.

M. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *LREC 2006*, 2006.

D. F. de Saussure. *Cours de linguistique generale*, chapter Cours de linguistique generale. Kluwer Academic Publishers, 1916.

M. Dong, K. Lua, and J. Xu. Selecting prosody parameters for unit selection based Chinese TTS. In K. Su, J. Tsujii, J. Lee, and O. Y. Kwong, editors, *Natural Language Processing – IJCNLP 2004*, volume 3248 of *Lecture Notes in Computer Science*, pages 272–279. Springer Berlin / Heidelberg, 2005.

M. Elsabrouty. DMET 1003 lecture: Automatic speech recognition, 2006. `http://cs.guc.edu.eg/courses/_Spring2008/DMET1003/slides/lecture12.pdf`.

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*, pages 417–422, 2006.

C. Fellbaum, editor. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press, 1999.

C. Féry. Syntax, information structure, embedded prosodic phrasing and the relational scaling of pitch accents. In N. Erteschick-Shir and L. Rochman, editors, *The sound of Syntax*, pages 271–290. Oxford University Press, 2009.

C. Féry. Recursion in prosodic structure. *Phonological Studies*, 13, 2010.

C. Féry and S. Ishihara. How focus and givenness shape prosody. In M. Zimmermann and C. Féry, editors, *Information Structure from Different Perspectives*, pages 36–63. Oxford University Press, 2009.

B. Fossett and P. Mirenda. *Handbook of Developmental Disabilities*, chapter Augmentative and Alternative Communication, pages 330–366. Guilford Press, 2009.

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

C. Gussenhoven. Focus, mode and the nucleus. *Journal of Linguistics*, 19:377–417, 1983.

C. Gussenhoven. *Thematic structure. Its role in grammar*, chapter Sentence accents and argument structure. Berlin, New York: Foris, 1992.

D. J. Hirst. Form and function in the representation of speech prosody. In K. Hirose, D. J. Hirst, and Y. Sagisaka, editors, *Quantitative prosody modeling for natural speech description and generation*, volume 46 of *Speech Communication*, pages 334–347. Elsevier, 2005.

D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Pearson Education, second edition, 2009.

L. Karttunen. Presuppositions and linguistic context. *Theoretical Linguistics*, 1:181–194, 1974.

K. Kipper, H. T. Dang, and M. Palmer. Class based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, Austin, Texas, 2000.

D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430, 2003.

B. Kollmeier, T. Brand, and B. Meyer. *Springer Handbook of Speech Processing*, chapter Perception of Speech and Sound, pages 61–82. Springer Berlin Heidelberg, 2008.

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

M. Krifka. Basic notions of information structure. In C. Féry, G. Fanselow, and M. Krifka, editors, *The Notions of Information Structure*, pages 13–55. Universitätsverlag Potsdam, 2007.

I. Kruijff-Korbayová and M. Steedman. Discourse and information structure. *Journal of Logic, Language and Information*, 12:249–259, 2003.

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.

M. Liberman and J. Pierrehumbert. Intonational invariance under changes in pitch range and length. In M. Aronoff and R.T. Oehrle, editors, *Language sound structure*, pages 157–233. Cambridge: MIT Press, 1984.

M. Liberman and A. Prince. On stress and linguistic rhythm. *Linguistic Inquiry*, 8:249–336, 1977.

E. D. Liddy. *Encyclopedia of Library and Information Science*, chapter Natural Language Processing. Marcel Dekker, second edition, 2001.

H. Liu and P. Singh. ConceptNet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22 (4):211–226, 2004.

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)*, pages 165–168, 2008.

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.

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

A. Neviarouskaya. *Compositional Approach for Automatic Recognition of Fine-Grained Affect, Judgment, and Appreciation in Text*. PhD thesis, University of Tokyo, 2010.

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

R. Picard. *Affective Computing*. Cambridge, MA: The MIT Press, 1997.

J. Pierrehumbert. *The phonology and phonetics of English intonation*. PhD thesis, MIT, 1980.

B. Z. Pollermann and M. Archinard. Improvements in speech synthesis. In E. Keller, G. Bailly, A. Monaghan, J. Terken, and M. Huckvale, editors, *Acoustic patterns of emotions*, pages 237–245. 2002.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

H. Quené. On the just noticeable difference for tempo in speech. *Journal of Phonetics*, 35(3):353–362, 2007.

G. I. Schlünz. The effects of part-of-speech tagging on text-to-speech synthesis for resource-scarce languages. Master's thesis, North-West University, Vanderbijlpark, South Africa, 2010.

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)*, pages 257–262, 2010.

M. Schröder. Expressive speech synthesis: Past, present, and possible futures. In J. Tao and T. Tan, editors, *Affective Information Processing*, pages 111–126. Springer London, 2009.

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

E. Selkirk. On derived domains in sentence phonology. *Phonology Yearbook*, 3:371–405, 1986.

E. Selkirk. *The handbook of phonological theory*, chapter Sentence prosody: intonation, stress and phrasing, pages 550–569. London: Blackwell, 1995.

E. Selkirk. The interaction of constraints on prosodic phrasing. In M. Horne, editor, *Prosody: Theory and Experiments*. Kluwer, 2000.

E. Selkirk. Contrastive focus, givenness and the unmarked status of "discourse new". In C. Féry, G. Fanselow, and M. Krifka, editors, *The Notions of Information Structure*, pages 125–145. Universitätsverlag Potsdam, 2007.

M. Shaikh, K. 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 the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, pages 326–329, Brisbane, Australia, 2008a.

M. Shaikh, H. Prendinger, and M. Ishizuka. Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. *Applied Artificial Intelligence*, 22:558–601, 2008b.

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

M. Shaikh, A. Rebordao, K. Hirose, and M. Ishizuka. Emotional speech synthesis by sensing affective information from text. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE, 2009b.

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.

J. O. Smith. Physical audio signal processing: Formant synthesis models. Online Book, December 2008a. `http://ccrma.stanford.edu/~jos/pasp/Formant_Synthesis_Models.html`.

J. O. Smith. Physical audio signal processing: Fundamental frequency estimation. Online Book, December 2008b. `http://ccrma.stanford.edu/~jos/pasp/Fundamental_Frequency_Estimation.html`.

R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. Normalization of non-standard words. Article submitted to Computer Speech and Language, 2001.

R. Stalnaker. Pragmatic presuppositions. In M. K. Munitz and P. K. Unger, editors, *Semantics and Philosophy*, pages 197–214. New York University Press, 1974.

M. Steedman. Structure and intonation. *Language*, 67:262–296, 1991.

M. Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 34:649–689, 2000.

M. Steedman. Information-structural semantics for english intonation. In C. Lee, M. Gordon, and D. Büring, editors, *Topic and focus: cross-linguistic perspectives on meaning and intonation*, pages 245–264. Springer, 2007.

B. R. Steunebrink. *The Logical Structure of Emotions*. PhD thesis, Utrecht University, The Netherlands, 2010.

C. Strapparava and A. Valitutti. WordNet-Affect: an affective extension of WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1083–1086, 2004.

T. Styger and E. Keller. *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*, chapter Formant synthesis, pages 109–128. John Wiley and Sons, Inc., 1994.

P. Taylor. *A Phonetic Model of English Intonation*. PhD thesis, University of Edinburgh, 1992.

P. Taylor. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*, 107(4):1697–1714, 2000.

P. Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, first edition, 2009.

K. Toutanova, D. Klein, C. D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 252–259, 2003.

H. Truckenbrodt. On the relation between syntactic phrases and phonological phrases. *Linguistic Inquiry*, 30(2):219–255, 1999.

H. Truckenbrodt. *The Encyclopedia of Languages and Linguistics*, volume 9, chapter Phrasal stress, pages 572–579. Oxford: Elsevier, second edition, 2006.

C. Tseng. Beyond sentence prosody. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, pages 20–29, 2010.

C. J. van Heerden, F. de Wet, and M. H. Davel. Automatic alignment of audiobooks in Afrikaans. In *Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, pages 187–191, 2012.

A. Waibel. *Prosody and Speech Recognition*. Pitman Publishing, London, first edition, 1988.

T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, 2005.

D. Yarowsky. Homograph disambiguation in text-to-speech synthesis. *Computer Speech and Language*, 1996.

S. J. Young, G. Evermann, M. J. F. Gales, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, Cambridge, UK, 2006.

H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 129–136, 2003.

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*, pages 294–299, 2007.

Appendix A

Tables of Significance

A.1 *t* Statistic

Table A.1: *t*-test table

df	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	3.078	6.314	12.706	31.821	63.657	318.313
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.782
8	1.397	1.860	2.306	2.896	3.355	4.499
9	1.383	1.833	2.262	2.821	3.250	4.296
10	1.372	1.812	2.228	2.764	3.169	4.143
11	1.363	1.796	2.201	2.718	3.106	4.024
12	1.356	1.782	2.179	2.681	3.055	3.929
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
70	1.294	1.667	1.994	2.381	2.648	3.211
80	1.292	1.664	1.990	2.374	2.639	3.195
90	1.291	1.662	1.987	2.368	2.632	3.183
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090

A.2 Chi-square Statistic

Table A.2: Chi-square table

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	–	–	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169