Semantics and Discourse Processing for Expressive TTS

Rodolfo Delmonte, Rocco Tripodi
Department of Linguistic Studies & Department of Computer Science
Ca' Foscari University of Venice
Email: delmont@unive.it

Abstract

In this paper we present ongoing work to produce an expressive TTS reader that can be used both in text and dialogue applications. The system has been previously used to read (English) poetry and it has now been extended to apply to short stories. The text is fully analyzed both at phonetic and phonological level, and at syntactic and semantic level. The core of the system is the Prosodic Manager which takes as input discourse structures and relations and uses this information to modify parameters for the TTS accordingly. The text is transformed into a poem-like structures, where each line corresponds to a Breath Group, semantically and syntactically consistent. Stanzas correspond to paragraph boundaries. Analogical parameters are related to ToBI theoretical indices but their number is doubled.

1 Introduction

In this paper we present ongoing work to produce an expressive TTS reader that can be used both in text and dialogue applications. The system has been previously used to read (English) poetry and we now decided to apply it to short stories. The text is fully analyzed both at phonetic and phonological level, and at syntactic and semantic level. In addition, the system has access to a restricted list of typical pragmatically marked phrases and expressions that are used to convey specific discourse function and speech acts and need specialized intonational contours.

Current TTS systems are dull and boring and characterized by a total lack of expressivity. They only take into account information coming from punctuation and in some cases, from tagging and syntactic constituency. Few expressive synthetic speech synthesizers are tuned to specific domains and are unable to generalize. They usually convey specific emotional content linked to a list of phrases or short utterances – see below. In particular, comma is a highly ambiguous punctuation mark with a whole set of different functions which are associated with specific intonational contours. In general, question and exclamative marks are used to modify the prosody of the previous word. We use the word "expressivity" in a specific general manner which includes sensible and sensitive reading that can only be achieved once a complete syntactic and semantic analysis has been provided to the TTS prosodic manager.

From a general point of view, the scientific problem can be framed inside the need to develop models that are predictive for a speech synthesizer to be able to sound natural and expressive, getting as close as possible to human-like performance. This can only be achieved manipulating prosody so that the text read aloud sounds fully natural, informative and engaging or convincing. However, in order to achieve something closer to that, text understanding should be attained or some similar higher level semantic computation. As Xu(2011) puts it, " It will probably be a long time before anything close to that is developed, of course"(ibid:94). Similar skeptical or totally negative opinions are expressed by Marc Huckvale (2002), when summarizing work he and his group have been carrying out for a long period over the project for an articulatory TTS called ProSynth. The goal of speech synthesis, in his perspective would be that of "understanding how humans talk" rather than the one of replicating a human talker (ibid. 1261).

Linguistically based work on emotions has been documented by the group working at Loquendo (now acquired by Nuance). They report their approach based on the selection of Expression which is related to a small inventory of what they call "speech acts" which coincide partly with dialogue, conversational and argumentative categories (Zovato et al. 2008; see also Campbell, 2002; Hamza et al. 2004). They implemented the acoustic counterpart of a limited, but rich, set of such categories, including: refuse, approval/ disapproval, recall in proximity, announce, request of information, request of confirmation, request of action/ behaviour, prohibition, contrast, disbelief, surprise/astonishment, regret, thanks, greetings, apologies, and compliments. In total, they

Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 32–43,
Lisboa, Portugal, 18 September 2015. ©2015 Association for Computational Linguistics.

managed to label and model accordingly some 500 different (expressive) utterances that can be used domain and context independently.

Work related to what we are trying to do is to be found in the field of storytelling and in experiments by the group from Columbia University working at MAGIC a system for the generation of medical reports. Montaño et al. [1] present an analysis of storytelling discourse modes and narrative situations, highlighting the great variability of speech modes characterized by changes in rhythm, pause lengths, variation of pitch and intensity and adding emotion to the voice in specific situations.

However, the approach most closely related to ours is the one by the group of researchers from Columbia University, where we can find Julia Hirschberg and Kathy McKeown. In the paper by S.Pan,K.McKeown & J.Hirschberg they highlight the main objectives of their current work, as "Prosody modeling" which is the task of "associating variations of prosodic features with changes in structure, meaning, intent and context of the language spoken." This requires "identifying correlations between this information and prosodic parameters through data exploration, and using learning algorithms to build prosody models from these data."(ibid. 1419) In fact, their attempt at using machine learning for prosody modeling has been only partially achieved. In their work on the concept-to-speech manager "the content planner uses a presentation strategy to determine and order content. It represents discourse structure, which is a hierarchical topic structure in MAGIC, discourse relations, which can be rhetorical relations, and discourse status, which represents whether a discourse entity is given, new or inferable and whether the entity is in contrast with another discourse entity."(ibid. 1420) As the authors affirm further on, the discourse level is where prosody is mostly affected. They then report previous work on discourse structure which can affect pitch range, pause and speaking rate by Grosz & Hirschberg, 1992; given/new/inferable information can affect pitch-accent placement by Hirschberg 1993; a shift in discourse focus can affect pitch-accent assignment (in Nakatani 1998); and contrastive entities can bear a special pitch accent (Prevost 1995). Further work towards predicting prosodic structure was published by Bachenko & Fitzpatrick, 1990, Delmonte & Dolci, 1991, and Wang & Hirschberg, 1992.

The objective of their experiment was modeling ToBI prosody features, i.e. pitch accents, phrase accents, boundary tones and break indices. Given the fact that there are six pitch-accent classes, five break-index classes, three phrase-accent classes, and three boundary-tone classes, they come up with a total of 17 different features organized in four separate classes. The experiment was carried out on a corpus of spontaneous speech with some 500 dialogues on medical issues, which ended up by being reduced to 250 annotated dialogues. In fact the features they managed to annotate are just surface syntactic and semantic ones[1].

The most disappointing fact was that they attempted to carry out a complete annotation but didn't succeed. In the paper they report their annotation efforts on the spontaneous-speech corpus which was automatically annotated with POS information, syntactic constituent boundaries, syntactic functions, and lexical repetitions, using approximations provided by POS taggers and parsers. It was also manually labelled with given/new/inferable information. But when it comes to semantic and discourse level information they say that they "are still working on manually labelling discourse structure, discourse relations, and semantic abnormality… We are currently annotating the speech corpus with features closely related to meaning and discourse."(ibid. 1426)

No further publication reports experiments with the complete annotation. And this is clearly due to difficulties inherent in the task. Now, this is what our system allows us to do, i.e. using discourse structure and relation to instruct the prosody manager to introduce the appropriate variation of prosodic parameters. According to ToBI features, this implies the ability to achieve: juncture placing prediction; phrase boundary tone prediction; prominence prediction; intonational contour movement prediction. To be more specific, given an input text the "Ideal

[1] and they are: (1) ID: the ID of a feature vector; (2) Lex: the word itself; (3) Concept: the semantic category of a content word; (4) SynFunc: the syntactic function of a word; (5) SemBoundary: the type of semantic constituent boundary after a word; (6) SemLength: the length, in number of words, of the semantic constituent associated with the current SemBoundary; (7) POS: the part-of-speech of a word; (8) IC: the semantic informativeness of a word(???), where in particular, the latter is – in our opinion – wrongly computed as a "semantic feature", being constitute by the logarithm of the relative frequency of a term in the corpus.

System" will read it aloud using naturally sounding prosody, where: phrasing is fully semantically consistent; intonation varies according to structural properties of clauses in discourse and speaker intention; prominence is assigned on the basis of novelty of topics and related events. In addition, expressivity conveys variations of attitude and mood as they are derived from deep subjective and affective analysis.

Our reformulation of ToBI (see Silverman et al. 1992) features from general/generic into concrete and implemented analogical parameters for natural and expressive TTS will be shown in a section at the end of the paper. The correspondence between prosodic features and linguistic representation is the issue to cope with and will be presented here.

Lieske et al.(1997) and Bos & Rupp(1998) documented their work on the generation system produced by the research project **Verb***mobil*. In particular the Verbmobil Interface Term which had responsibility for the interaction between different linguistic modules, including a TTS and an ASR modules. These linguistic modules included a SynSem component, i.e. a syntactic, a semantic and discourse component, which was meant to drive the generation of appropriate utterance with the appropriate prosody. The prosody component of Verbmobil is related to semantics and can influence segmentation, sentence mood and focus. The ad-hoc formalism they created allowed the parser to take into account prosodic information already from the start. However, the Verbmobil system did not allow to communicate stress patterns to the TTS. Here we are dealing with a much simpler effort which also has semantics and other discourse level information available from the generator. On the contrary, SPARSAR is a system that can be used with any English text or poem and has to derive its information directly from the words.

2 Semantic Representation for TTS

Systems that can produce an appropriate semantic representation for a TTS are not many at an international level but they can be traced from the results of a Shared Task organized by members of SigSem and are listed here below in the corresponding webpage http://www.sigsem.org/w/index.php?title=STEP_2008_shared_task:_comparing_semantic_representations (see Bos & Delmonte, 2008).

State of the art semantic systems are based on different theories and representations, but the final aim of the workshop was reaching a consensus on what constituted a reasonably complete semantic representation. Semantics in our case not only refers to predicate-argument structure, negation scope, quantified structures, anaphora resolution and other similar items, it refers essentially to a propositional level analysis. Propositional level semantic representation is the basis for discourse structure and discourse semantics contained in discourse relations. It also paves the way for a deep sentiment or affective analysis of every utterance, which alone can take into account the various contributions that may come from syntactic structures like NPs and APs where affectively marked words may be contained. Their contribution needs to be computed in a strictly compositional manner with respect to the meaning associated to the main verb, where negation may be lexically expressed or simply lexically incorporated in the verb meaning itself.

In Fig. 1 we show the architecture of our deep system for semantic and pragmatic processing, in which phonetics, prosodics and NLP are deeply interwoven.

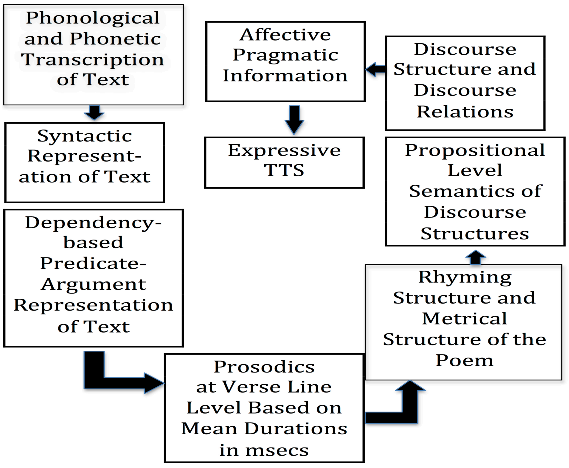

Phonological and Phonetic Transcription of Text
Syntactic Represent-ation of Text
Dependency-based Predicate-Argument Representation of Text
Affective Pragmatic Information
Expressive TTS
Discourse Structure and Discourse Relations
Propositional Level Semantics of Discourse Structures
Rhyming Structure and Metrical Structure of the Poem
Prosodics at Verse Line Level Based on Mean Durations in msecs

Figure 1. System Architecture Modules for SPARSAR

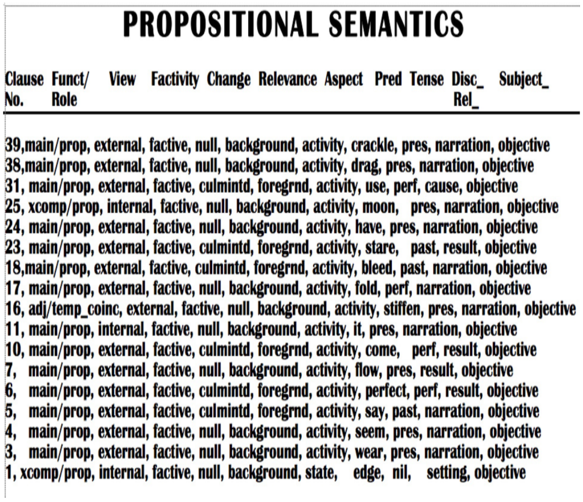
The system is based on VENSES a shallow version of GETARUNS. All these versions have been extensively tested and results published in a number of international publications and collected in two books (Delmonte 2007;2009)[1].

[1] In more detail here: Parser evaluation has been documented in Delmonte(2002;2004); Anaphora Resolution in Delmonte(2002a), Delmonte et al.(2006a), Delmonte & Tonelli(2006); Discourse Relations in

The current system may take any English text and produce an output to be used for TTS. All components of the system have undergone evaluation in particular discourse level analysis has been shown to be particularly effective (see Delmonte, 2007). The system does low level analyses before semantic modules are activated, that is tokenization, sentence splitting, multiword creation from a large lexical database. Then chunking and syntactic constituency parsing which is done using a rule-based recursive transition network. The parser works in a cascaded recursive way to include always higher syntactic structures up to sentence and complex sentence level. These structures are then passed to the first semantic mapping algorithm that looks for subcategorization frames in the lexica made available for English, including VerbNet, FrameNet, WordNet and a proprietor lexicon with most frequent verbs, adjectives and nouns, containing also a detailed classification of all grammatical or function words. This mapping is done following LFG principles, where c-structure is turned into f-structure obeying uniqueness, completeness and coherence grammatical principles. The output of this mapping is a rich dependency structure, which contains information related also to implicit arguments, i.e. subjects of infinitivals, participials and gerundives. It also has a semantic role associated to each grammatical function, that is used to identify the syntactic head lemma uniquely in the sentence. Finally it takes care of long distance dependencies for relative and interrogative clauses.

Now that fully coherent and complete predicate argument structures have been built, pronominal binding and anaphora resolution algorithms can be fired. Also coreferential processed are activated at the semantic level: they include a centering algorithm for topic instantiation and memorization that we do using a three-place stack containing a Main Topic, a Secondary Topic and an Potential Topic. In order to become a Main Topic, a Potential Topic must be reiterated and become persistent in the text.

Discourse Level computation is done at propositional level by building a vector of features associated to the main verb complex of each clause. They include information about tense, aspect, negation, adverbial modifiers, modality. These features are then filtered through a set of rules which have the task to classify a proposition as either objective/subjective, factual/nonfactual, foreground/background. In addition, every lexical predicate is evaluated with respect to a class of discourse relations. Eventually, discourse structure is built, according to criteria of clause dependency in which a clause can be classified either as coordinate or subordinate. As a result, we have a set of four different moves to associate to each clause: root, down, level, up. We report here below semantic and discourse structures related to the poem by Sylvia Plath "Edge" which you can find here, http://www.poetryfoundation.org/poem/178970.



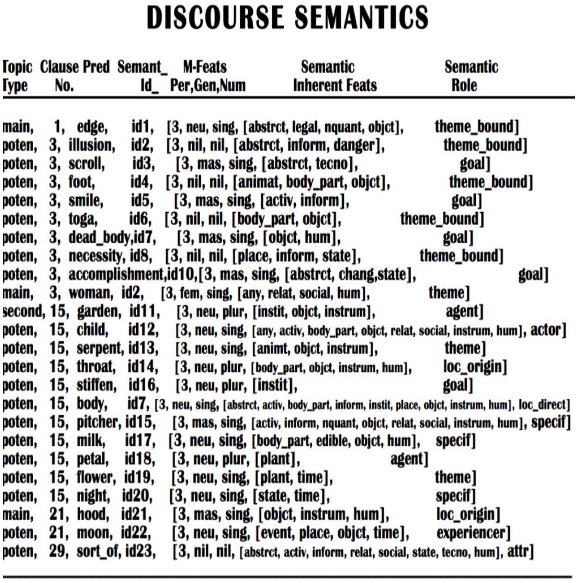
```
PROPOSITIONAL SEMANTICS

Clause Funct/   View   Factivity Change Relevance Aspect  Pred Tense Disc_   Subject_
No.    Role                                                          Rel_

39,main/prop, external, factive, null, background, activity, crackle, pres, narration, objective
38,main/prop, external, factive, null, background, activity, drag, pres, narration, objective
31, main/prop, external, factive, culmintd, foregrnd, activity, use, perf, cause, objective
25, xcomp/prop, internal, factive, null, background, activity, moon,   pres, narration, objective
24, main/prop, external, factive, null, background, activity, have, pres, narration, objective
23, main/prop, external, factive, culmintd, foregrnd, activity, stare,   past, result, objective
18,main/prop, external, factive, culmintd, foregrnd, activity, bleed, past, narration, objective
17, main/prop, external, factive, null, background, activity, fold, perf, narration, objective
16, adj/temp_coinc, external, factive, null, background, activity, stiffen, pres, narration, objective
11, main/prop, internal, factive, null, background, activity, it, pres, narration, objective
10, main/prop, external, factive, culmintd, foregrnd, activity, come,   perf, result, objective
7,   main/prop, external, factive, null, background, activity, flow, pres, result, objective
6,   main/prop, external, factive, culmintd, foregrnd, activity, perfect, perf, result, objective
5,   main/prop, external, factive, culmintd, foregrnd, activity, say, past, narration, objective
4,   main/prop, external, factive, null, background, activity, seem, pres, narration, objective
3,   main/prop, external, factive, null, background, activity, wear, pres, narration, objective
1, xcomp/prop, internal, factive, null, background, state,   edge,  nil,   setting, objective
```

Figure 2. Propositional semantics for Edge

In Fig.2, clauses governed by a copulative verb like BE report the content of the predication to the subject. The feature CHANGE can either be set to NULL, GRADED or CULMINATED: in this case Graded is not used seen that there no progressive or overlapping events.

In the representation of Figure.3, we see topics of discourse as they have been computed by the coreference algorithm, using semantic indices characterized by identifiers starting with ID. Every topic is associated to a label coming from the centering algorithm: in particular, WOMAN which is assigned ID id2 reappears as MAIN topic in clauses marked by no. 15. Also BODY reappears with id7. Every topic is associated to

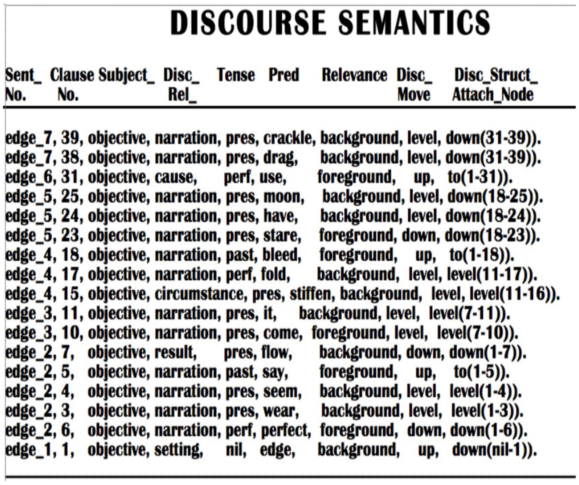
Delmonte et al.(2007a;2007b); Recognizing Textual Entailment evaluations in Delmonte et al.(2005), Delmonte et al.(2006b), Delmonte, Bristot, Piccolino, Tonelli, (2007), Delmonte et al.(2009); Implicit Entities and Events in Delmonte & Pianta(2009), Delmonte(2009a;b;c), Delmonte & Tonelli(2010); Tonelli & Delmonte(2011); Delmonte(2013) morphological features, semantic inherent features and a semantic role.



```
                DISCOURSE SEMANTICS

Topic Clause Pred Semant_  M-Feats          Semantic           Semantic
Type   No.          Id_  Per,Gen,Num      Inherent Feats         Role

main,   1,  edge,     id1,  [3, neu, sing, [abstrct, legal, nquant, objct],     theme_bound]
poten,  3,  illusion,  id2,  [3, nil, nil, [abstrct, inform, danger],           theme_bound]
poten,  3,  scroll,    id3,   [3, mas, sing, [abstrct, tecno],                   goal]
poten,  3,  foot,      id4,  [3, nil, nil, [animat, body_part, objct],           theme_bound]
poten,  3,  smile,     id5,   [3, mas, sing, [activ, inform],                    goal]
poten,  3,  toga,      id6,  [3, nil, nil, [body_part, objct],            theme_bound]
poten,  3,  dead_body,id7,    [3, mas, sing, [objct, hum],                      goal]
poten,  3,  necessity, id8,  [3, nil, nil, [place, inform, state],          theme_bound]
poten,  3,  accomplishment,id10,[3, mas, sing, [abstrct, chang,state],                     goal]
main,   3,  woman,  id2,   [3, fem, sing, [any, relat, social, hum],          theme]
second, 15, garden, id11,   [3, neu, plur, [instit, objct, instrum],              agent]
poten,  15, child,    id12,   [3, neu, sing, [any, activ, body_part, objct, relat, social, instrum, hum], actor]
poten,  15, serpent, id13,   [3, neu, sing, [animt, objct, instrum],              theme]
poten,  15, throat,  id14,   [3, neu, plur, [body_part, objct, instrum, hum],      loc_origin]
poten,  15, stiffen,  id16,   [3, neu, plur, [instit],                            goal]
poten,  15, body,    id7, [3, neu, sing, [abstrct, activ, body_part, inform, instit, place, objct, instrum, hum], loc_direct]
poten,  15, pitcher, id15,   [3, mas, sing, [activ, inform, nquant, objct, relat, social, instrum, hum], specif]
poten,  15, milk,    id17,  [3, neu, sing, [body_part, edible, objct, hum],   specif]
poten,  15, petal,   id18,   [3, neu, plur, [plant],                   agent]
poten,  15, flower, id19,   [3, neu, sing, [plant, time],                      theme]
poten,  15, night,  id20,   [3, neu, sing, [state, time],                      specif]
main,   21, hood,   id21,    [3, mas, sing, [objct, instrum, hum],             loc_origin]
poten,  21, moon,  id22,    [3, neu, sing, [event, place, objct, time],        experiencer]
poten,  29, sort_of, id23,   [3, nil, nil, [abstrct, activ, inform, relat, social, state, tecno, hum], attr]
```

Figure 3. Discourse level Semantics for Topic Hierarchy

Eventually, the final computation concerning Discourse Structure is this one:



```
                 DISCOURSE SEMANTICS

Sent_ Clause Subject_ Disc_   Tense  Pred    Relevance Disc_   Disc_Struct_
No.    No.             Rel_                            Move   Attach_Node

edge_7, 39, objective, narration, pres, crackle, background, level, down(31-39)).
edge_7, 38, objective, narration, pres, drag,    background, level, down(31-39)).
edge_6, 31, objective, cause,     perf, use,     foreground,   up,  to(1-31)).
edge_5, 25, objective, narration, pres, moon,    background, level, down(18-25)).
edge_5, 24, objective, narration, pres, have,    background, level, down(18-24)).
edge_5, 23, objective, narration, pres, stare,   foreground, down, down(18-23)).
edge_4, 18, objective, narration, past, bleed,   foreground,   up,  to(1-18)).
edge_4, 17, objective, narration, perf, fold,    background,  level, level(11-17)).
edge_4, 15, objective, circumstance, pres, stiffen, background,  level, level(11-16)).
edge_3, 11, objective, narration, pres, it,    background, level,  level(7-11)).
edge_3, 10, objective, narration, pres, come,  foreground, level,  level(7-10)).
edge_2, 7,  objective, result,    pres, flow,    background, down, down(1-7)).
edge_2, 5,  objective, narration, past, say,     foreground,   up,   to(1-5)).
edge_2, 4,  objective, narration, pres, seem,    background, level,  level(1-4)).
edge_2, 3,  objective, narration, pres, wear,    background, level,  level(1-3)).
edge_2, 6,  objective, narration, perf, perfect, foreground,  down, down(1-6)).
edge_1, 1,  objective, setting,    nil,  edge,    background,   up,  down(nil-1)).
```

Figure 4. Discourse Semantics for Discourse Structures

Movements in the intonational contours are predicted to take place when FOREGROUND and UP moves are present in the features associated to each clause.

2.1 From Poetry to Story Reading

We referred to a poem in the previous section because in fact we will be using rules associated to poetry prosodic mapping in our work on story reading. We assume that reading a story aloud requires the reader to organize pauses in such a way that expressivity and meaning is preserved. This process is usually referred to as Breath Group organization. Since a breath group is a well-formed group of words conveying a concept or a meaning we decided to compare it to a line in a poem. Poems are organized into lines and stanzas, while stories usually have punctuation to mark main concepts and introduce pauses. Punctuation however is not sufficient in itself and does not always guarantee meaning coherence. In particular, Commas are highly ambiguous and may be used for a whole set of different functions in discourse. So eventually what we can actually trust are Breath Groups. Continuing our comparison with poems, lines may be end-stopped or enjambed when they run on the following line or stanza. The same may happen with Breath Groups, they may be end-stopped or enjambed and require a different prosodic setup.

We will then define Breath Groups as syntactically and semantically coherent units coinciding with an Intonation Phrase in ToBI terms: IPs are characterized by different tones, possible boundary tones and break indices. On the contrary, pitch Accents are associated to word stresses which are present in our phonetic representation: except that only syntactic heads are associated with Pitch Accents, dependents are demoted.

2.2 Implementing the Rules for Expressive TTS

Let's now look at one example, a short story by Aesop, "Bellying the Cat" that can be found here, http://www.taleswithmorals.com/aesop-fable-belling-the-cat.htm. At first we show the decomposition of the story into Breath Groups and then the mapping done by the Prosodic Manager.

long_ago ß
the mice had a general council ß
to consider what measures they could take ß
to outwit their common enemy ß
the cat ß
some said this ß
and some said that ß
but at_last a young mouse got_up ß
and said he had a proposal ß
to make ß
which he thought would meet the case ßß
you will all agree ß
said he ß
that our chief danger consists in the sly ß
and treacherous manner ß
in which the enemy approaches us ßß
now ß

if we could receive some signal of her approach ß
we could easily escape from her ß
i venture ß
therefore ß
to propose that a small bell be procured ß
and attached by a ribbon round the neck of the cat ß
by_this_means ß
we should always know when she was about ß
and could easily retire ß
while she was in the neighborhood ßß
this proposal met with general applause ß
until an old mouse got_up ß
and said ßß
that is all very_well ß
but who is to bell the cat ßß
the mice looked at one_another ß
and nobody spoke ß
then the old mouse said ßß
it is easy ß
to propose impossible remedies ßß

Table 1. Decomposition of the text into Breath Groups

2.3 Breath Group Creation Rules

A first set of the rules to map the story into this structures are reported below. The rules are organized into two separate sets: low level and high level rules. Here are low level ones:
- Follow punctuation first, but check constituent length; look for Coordinate Structures;
- look for Subordinate Clauses;
- look for Infinitival Complements;
- look for Complement Clauses; look for Relative Clauses;
- look for Subject and VerbPhrase juncture;
- look for AdverbialPhrase but only when beginning of Clause;
- look for Obligatory complements followed by adjuncts - with long constituents (Constituent length is at first checked in no. of words but also by phonetic length in no. of phones and their average duration in msec).

The high level corresponds to the recursive level. Recursive rules are associated with complex sentences and with Coordinate, Subordinate and Complement clauses. In Appendix 1 we show the mapping into Analogical phonetic acoustic correlates of pitch, speaking rate and intensity, and pauses for the text above. They can be copy/pasted into a TextEdit file and spoken aloud by Apple TTS.

3 The Prosodic Manager or ToBI features re-implemented

We will now discuss the use of Pierrehumbert's inventory of Tones and Break Indices, in relation to its actual application in real texts reading. We shall start from Break Indices which amount to 5, starting from 0 to 4 included. We assume that BI 0 is in a sense very special and peculiar and covers an aspect of prosody which has no general and practical application. As for BI 2 we will use label it to cover one of the phenomena indicated in the manual, that the idea to indicate a stronger sense of disjuncture than 1, for careful deliberation (see manual online).

So we come up with two types of BIs: those that are simple pauses, and those that also induce an intonation curve reset. BI 3 and 4 are intonationally related and regard phrase and sentence level prosody. BI 1 and B 2 are to be regarded as pertaining to word level and to possible internal constituent or internal phrase. The latter BIs have no effect on the intonational contour. In terms of our analogical parameterization, the former two indices require a ***reset*** at the end that accompany the ***silence***, the latter two have no ***reset***. However, our list is much longer:

[[slnc 300]],[[rset 0]]	**BI 4**
[[slnc 200]],[[rset 0]]	**BI 3**
[[slnc 100]]	**BI 2**
[[slnc 30]],[[rset 0]]	**BI 32**
[[slnc 50]],[[rset 0]]	**BI 33**
[[slnc 100]],[[rset 0]]	**BI 23**
[[slnc 300]]	**BI 22**
[[slnc 400]]	**BI 44**
[[rate 110; volm +0.3]]	**<slow down**
[[rate 130; volm +0.5]]	**<slow down**

In our representation, there are additional different 2 and 3 breaks: the reason for that is due to the use of the break in presence of end of Breath Group, with punctuation (BI 3) and without punctuation. The latter case is then split into two subcases, one in which the last word – a syntactic head – is followed or not by a dependent word, hence 33 and 32 respectively are the indices used. We also use 44 for the separation of the title from the text. Finally 23 is a break with a reset between constituents of a specific type, quantifiers. Then we have two slow down commands, one that precedes again quantifiers, and the other for all syntactic heads, end of Breath Groups (hence BGs). Quantifiers are treated in a special manner by the system if they are syntactic heads. For instance consider "Nothing" which is a subject head,

[[rate 110; volm +0.3]]	**<slow down**
[[slnc 100]],[[rset 0]]	**BI 2 %**

Coming now to tones and pitch accents, we assume the original list is again insufficient to cover the phenomena we found in our texts. We show the list of additional labels in Table 2 in Appendix 2.

3.1 The algoritm of the Prosodic Manager

The algorithm of the Prosodic Manager (hence PM) is a continuation of work carried out by Delmonte (1985). It receives information from the syntactic level - all heads with their grammatical function; from the semantic level - all discourse relations and structures with their relevance function foreground/background; from the metrical level - all end of line (Breath Groups) words with their relative line number plus all end of stanza lines again with relative line number; all phonetically translated words at each line level. And of course all sentences into which the text has been automatically split.

The PM receives one sentence at a time from the list of sentences and passes it down to the recursive algorithm that has the task of transforming all these rules into analogical parameters for the TTS. The first sentence coincides with the title and author and is computed in a standardized way. The computation starts from the first line of the first stanza: now the PM has to match the information available at sentence level with the subdivision of the text into lines or BGs. Sentences do not coincide with lines nor with stanzas. In some cases, when lines are end-stopped with a period as punctuation it may be the case that they coincide with a single sentence. However this is usually rare. Three indices are then needed to keep trace of what the recursive algorithm is doing and where in the text it is positioned. This is due to the fact that the end of line position may contain words that may occur in multiple places, both at the end and line internally. In order to help with recognizing where the PM is positioned, we collect all stanza markers with their indices, taken from the list of end of line last words.

So, the PM keeps note of each word in a sentence with an internal index; it then keeps note of the end of stanza by removing stanza markers both in the list of end of line words and in the list of end of stanzas. The input string is the one coming from the list of words contained in the sentence. When we meet a word which is recognized as end of line, we then check to see whether this word is followed by punctuation or not. In case it is not followed by punctuation we check to see whether the rest of the sentence contains other identical words and whether these are end of line. If the current word is not present in the rest of the sentence then is last, else if it is present more than once in the list of LAST words again is last. Now the system knows at what Stanza it is positioned and can verify whether the current word matches the last of the current stanza. To do this, we find all the N-stanzas that have the N lower than the index associated to the LAST word found - this should match with the current stanza number.

The PM has 35 high level recursive rules, these in turn contain the following associated rules:

- discourse level rules: removeforeground ; removebackground five different calls. They fire a specific intonational control parametric combination, for FOREGROUND discourse structures, and for BACKGROUND discourse structures;
- discourse level rules: direct speech is fired by a first sentence and is then continued in one or more following sentence/s thus requiring the downstep intonation to be in place. This has to continue until the final sentence of direct speech is detected. If downstep was not in place, the sentence would be computed with a normal reset and a possible declarative simple declination line with no relation whatsoever with the previous sentence in discourse.
- syntactic-phonological rules : these rules check to see whether the current word or the pair of current words are end-of-line and if yes whether they are syntactic heads or not ;
- this will trigger the parameter [[rate 130; volm +0.5]] for BI and possible boundary tones depending on position with respect to stanza ending;
- rules for multiwords are needed to restructure these words and see whether they are part of the list of affective words ;
- rules for affective words and phrases : they have to be treated differently according to whether they are heads or dependents, line final or not; they are associated with a descending tone;
- semantic rules devised for exclamatives and questions, their tone is raised and the speaking rate is also raised;
- exceptions rules: these have been created to account for the role of specific items in the sentence which have been previously computed like discourse markers introducing coordination and comparisons; a short list of conjunction with a concessive or adversative content. Finally a list

which contains words that Apple TTS cannot pronounce correctly and need our phonetic reconversion. Then just exceptions constituted by quantifiers which have been computed as syntactic heads and require to be set apart by introducing a specific BI.

- pragmatic rules : for lexically frozen expression and for particular emotionally and conversationally related phrases and utterances. These have been organized as rules to modify the phrase or utterance, depending on the specific dialogue act, emotion or conversational turn it refers to subdividing the tone sequence possibily in bitonal pitch accent.

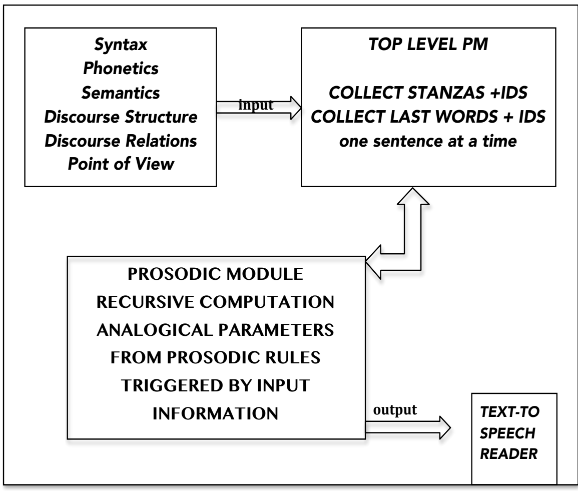


Figure 5: The Algorithm of the Prosodic Manager

3.2 One specific case: downstepped Direct Speech

Consider now the case of another of the fables by Aesop we worked on – The Fox and the Crow, that can be found here, http://www.taleswithmorals.com/aesop-fable-the-fox-and-the-crow.htm. In this story the main character introduced by the narrator, starts his speech with an exclamative sentence and then continues with some explanation and elaborations. These discourse structures need to be connected to the previous level of intonation. This requires receiving information at clause level from the discourse level, in order to allow for the appropriate continuation. In particular, this is done by:

- detecting the presence of Direct Speech by both verifying the presence of a communication verb governor of a sentence started by the appropriate punctuation mark, inverted commas. This same marker will have to be detected at the end of direct speech. The end may coincide with current sentence or a number of additional sentences might be present as is the case at stake. The current reported speaker continues after the exclamative with a series of apparently neutral declarative sentences, which can be computed as explanations and elaborations. But they all depend from the exclamative and need to be treated accordingly at discourse level.

To work at discourse level, the system has a switch called “point of view” which takes into account whose point of view is reported in each sentence. The default value for a narrative text would be the “narrator” whenever the sentence is reported directly with no attribution of what is being said. When direct speech appears, the point of view is switched to the character whom the sentence has been attributed to. This switch is maintained until the appropriate punctuation mark appears. So eventually, it is sufficient for the PM to take the current point_of_view under control. If it is identical to the previous one, nothing happens. If it changes to a new holder and it is marked with direct speech, the algorithm will be transferred to a different recursive procedure which will continue until point_of_view remains identical. This new procedure allows the PM to assign downstepped intonational contours as shown here below. In this fragment, we also mark the presence of a word – HUE - which is wrongly pronounced by Apple synthesizer and requires activating the exceptional phonetic conversion.

"What a noble bird I see **BI-3** above me **BI-22 H*-H-1** ! **BI-2 H-!H*-1**
Her beauty is without **H*-L%** equal **BI-3** ,
H*-L the [[inpt PHON]]hUW[[inpt TEXT]] of her plumage **H*-H-3** exquisite **BI-2 .**
H-!H*-1 If only her voice is **BI-2** as sweet **BI-2** as her **BI-2 H-!H*-1** looks are **H*-L** fair **BI-3** ,
she **BI-2 H-H*-2** ought **L*-L%** without doubt [[rset 0]] to be Queen of the **H*-L%-2** Birds **BI-3** . “

In case this information was not made available to the PM, the result would have been the following.

" What a noble bird I see **BI-3** above me **BI-22 H*-H-1** ! **BI-2 H-!H*-1**!
Her beauty is without **H*-L%** equal **BI-3** ,

H*-L the [[inpt PHON]]hUW[[inpt TEXT]] of her plumage **H*-H-3** exquisite **BI-2** .
If only her voice is **BI-2** as sweet **BI-2**
as her **BI-2 H-!H*-1** looks are **H*-L** fair **BI-3** ,
she **BI-2 H-H*-2** ought **L*-L%** without doubt [[rset 0]] to be Queen of the **H*-L%-2** Birds **BI-3** . "

We started lately to experiment with Google Chrome addon TTS which is included in the system SpeakIt© and contains iSpell TTS. Some of the voices are particularly well equipped and we tested English UK female. The TTS requires a fee to be paid and the use of an XML interface based on SSML, Speech Synthesis Markup Language adopted by W3C, Version 1.1. The authors of the specification unclude well-known experts of speech synthesis and prosody, like Paolo Baggia from Loquendo, Paul Bagshaw from France Telecom. The excerpt from Aesop's story converted into this new language is given here below. Note that the conversion has been done using the new ToBI labels:

```
What a noble <prosody pitch="medium">bird I see</prosody><break time="100ms"/>
<prosody pitch="default" rate="slow" volume="-0.2">above me </prosody><break time="200ms"/>
<prosody pitch="medium" rate="medium" volume="+1.1">Her beauty is without </prosody> <prosody pitch="-10Hz" rate="default" volume="medium"> equal ,</prosody> <break time="200ms"/>
<prosody rate="default" volume="medium">the hue of her plumage</prosody>
<prosody pitch="medium" rate="default" volume="+1.1">exquisite</prosody><break time="200ms"/>
<prosody pitch="low" rate="default" volume="loud">If only her voice </prosody><break time="5ms"/><prosody pitch="low" rate="default" volume="loud">is as sweet </prosody> <break time="10ms"/>
<prosody pitch="medium" rate="default" volume="loud">as her looks are fair</prosody><break time="200ms"/>
<prosody pitch="medium" rate="default" volume="medium">she <break time="5ms"/> </prosody><prosody pitch="high" rate="slow" volume="loud">ought</prosody><prosody pitch="medium" rate="slow" volume="soft">without doubt to be Queen of the</prosody>
<prosody pitch="high" rate="default" volume="loud">Birds</prosody></speak>
```

5 Evaluation and Conclusion

The system has undergone extensive auditory evaluation by expert linguists. It has also been presented at various demo sessions always receiving astounded favourable comments (Delmonte & Bacalu, 2013; Delmonte & Prati, 2014; Delmonte 2015). The evaluation has been organized in two phases, at first the story is read by Apple TTS directly from the text. Then the second reading has been done by the system and a comparison is asked of the subject listening to it. In the future we intend to produce an objective evaluation on a graded scale using naïve listeners English native speakers. We will be using the proposal in Xu (2011:95), called MOS, or Mean Opinion Score, with a five-level scale: 5-Excellent, 4-Good, 3-Fair, 2-Poor, 1-Bad, with the associated opinions: 5-Imperceptible, 4-Perceptible but not annoying, 4-Slightly annoying, 2-Annoying, 1-Very annoying.

In this paper we presented a prototype of a complete system for expressive and natural reading which is fully based on internal representations produced by syntactic and semantic deep analysis. The level of computation that is mostly responsible for prosodic variations is the discourse level, where both discourse relations, discourse structures, topic and temporal interpretation allow the system to set up an interwoven concatenation of parameters at complex clause and sentence level. Pragmatically frozen phrases and utterances are also fully taken into account always at a parameterized level. Parameters have been related to ToBI standard set and a new inventory has been proposed. The system is currently working on top of Apple TTS but we already started to port it to other platforms. It is available for free download at a dedicated website.

References

Balyan, Archana, S. S. Agrawal, Amita Dev, Speech Synthesis: A Review, International Journal of Engineering Research & Technology (IJERT), Vol. 2, Issue 6, 57-75.

Bachenko, J. & Fitzpatrick, E. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Comp. Ling.* 16, 155{170.

Bos, Johan and C.J. Rupp, Bianka Buschbeck-Wolf and Michael Dorna, Managing information at linguistic interfaces, 1998. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International

Conference on Computational Linguistics, Volume 1, ACL-COLING, pp. 160-166.

Bos Johan & Rodolfo Delmonte (eds.), 2008. Semantics in Text Processing (STEP), Research in Computational Semantics, Vol.1, College Publications, London.

Campbell N., 2002. Towards a grammar of spoken language: incorporating paralinguistic information. In: 7th ISCA International Conference on Spoken Language Processing, Denver, Colorado, USA, September 16-20, 2002.

Campbell's Conclusion for SSW6 TALK on Towards Conversational Speech Synthesis; "Lessons Learned from the Expressive Speech Processing Project".http://www.isca-speech.org/ archive_open/archive_papers/ssw6/material/ ssw6_022/ ssw6_022.ppt

Raúl Montaño, Francesc Alías, Josep Ferrer, 2013. Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis, 8th ISCA Speech Synthesis Workshop.

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

Delmonte R., 1982. Automatic Word-Stress Patterns Assignment by Rules: a Computer Program for Standard Italian, Proc. IV F.A.S.E. Symposium, 1, ESA, Roma, 153-156.

Delmonte R., G.A.Mian, G.Tisato, 1984. A Text-to-Speech System for the Synthesis of Italian, Proceedings of ICASSP'84, San Diego(Cal), 291-294.

Delmonte R., 1986. A Computational Model for a text-to-speech translator in Italian, Revue - Informatique et Statistique dans les Sciences humaines, XXII, 1-4, 23-65.

Delmonte R., G.A.Mian, G.Tisato, 1986. A Grammatical Component for a Text-to-Speech System, Proceedings of the ICASSP'86, IEEE, Tokyo, 2407-2410.

Delmonte R., R. Dolci, 1991. Computing Linguistic Knowledge for text-to-speech systems with PROSO, Proceedings 2nd European Conference on Speech Communication and Technology, Genova, ESCA, 1291-1294.

Delmonte R., 2002. GETARUN PARSER - A parser equipped with Quantifier Raising and Anaphoric Binding based on LFG, Proc. LFG2002 Conference, Athens, pp.130-153, at http://cslipublications.stanford.edu/hand/miscpubs online.html.

Delmonte R., 2002a. From Deep to Shallow Anaphora Resolution:, in Proc. DAARC2002 , 4th Discourse Anaphora and Anaphora Resolution Colloquium, Lisbona, pp.57-62.

Delmonte R., 2004. Evaluating GETARUNS Parser with GREVAL Test Suite, Proc. ROMAND - 20th International Conference on Computational Linguistics - COLING, University of Geneva, 32-41.

Delmonte R., S.Tonelli, M.A. Piccolino Boniforti, A. Bristot, E.Pianta, 2005. VENSES – a Linguistically-Based System for Semantic Evaluation, in Joaquin Quiñonero-Candela, et al., 2005, Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment.: MLCW 2005, UK, 344-371.

Delmonte R., A.Bristot, M.A. Piccolino Boniforti and S.Tonelli, 2006a. Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach, ROMAND 2006, 11th EACL, Trento, Association for Computational Linguistics, 3-10.

Delmonte R., A. Bristot, M.A. Piccolino Boniforti and S. Tonelli, 2006b. Coping with semantic uncertainty with VENSES, in Bernardo Magnini, Ido Dagan(eds.), Proceedings of the Challanges Workshop - The 2nd PASCAL Recognizing Textual Entailment Challenge, Università Ca' Foscari, Venezia, 86-91.

Delmonte R., A.Bristot, M.A.Piccolino Boniforti, S.Tonelli, 2007. Entailment and Anaphora Resolution in RTE3, in Proc. ACL Workshop on Text Entailment and Paraphrasing, Prague, ACL Madison, USA, pp. 48-53.

Delmonte R., G. Nicolae, S. Harabagiu (2007b), A Linguistically-based Approach to Detect Causality Relations in Unrestricted Text, in Proc. MICAI-2007, IEEE Publications, 173-185.

Delmonte R., G. Nicolae, S. Harabagiu, (2007a), A Linguistically-based Approach to Discourse Relations Recognition, in B.Sharp & M.Zock(eds.), Natural Language Processing and Cognitive Science, Proc. 4th NLPCS, Funchal, Portugal, INSTICC PRESS, pp. 81-91.

Delmonte, R, & S. Tonelli, 2009. Knowledge-poor and Knowledge-rich Approach in Anaphora Resolution Algorithms : a Comparison, In Linguistica e modelli tecnologici di ricerca : atti del XL Congresso internazionale di studi della Società di linguistica italiana (SLI), Roma, Bulzoni, pp.1-7.

Delmonte R., S.Tonelli, R. Tripodi, (2009), Semantic Processing for Text Entailment with VENSES, in Proceedings of Text Analysis Conference (TAC) 2009 Workshop - Notebook Papers and Results, NIST, Gaithersburg MA, pp. 453-460.

Delmonte R., E. Pianta, 2009. Computing Implicit Entities and Events for Story Understanding, in H.Bunt, V.Petukhova and S.Wubben(eds.), Proc. Eighth International Conference on Computational Semantics IWCS-8, Tilburg University Press, pp. 277-281.

Delmonte R., 2009a. Computing Implicit Entities and Events with Getaruns, in B.Sharp and M.Zock (eds.), Natural Language Processing and Cognitive Science 2009, Insticc Press, Portugal, 23-35.

Delmonte R., 2009b. A computational approach to implicit entities and events in text and discourse, in International Journal of Speech Technology (IJST), Springer, pp. 1-14.

Delmonte R., 2009c. "Understanding Implicit Entities and Events with Getaruns," IEEE International Conference on Semantic Computing, Berkeley, pp. 25-32.

Delmonte R. & S. Tonelli, 2010. VENSES++-UNIVE: Adapting a deep semantic processing system to the identification of null instantiations, Proceedings of the 5th International Workshop on Semantic Evaluation, ACL, pp. 296–299.

Tonelli S., R. Delmonte, 2011. "Desperately seeking Implicit arguments in text", in RELMS'2011, Workshop on Relational Models of Semantics at ACL 2011 Portland, pp.54-62.

Delmonte R., 2013. Coping With Implicit Arguments And Events Coreference, in E. Hovy, T. Mitamura, M. Palmer, (eds.), Proceedings of the Conference The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation, HLT-NAACL, Atlanta, pp. 1-10.

Delmonte R. & C. Bacalu. 2013. SPARSAR: a System for Poetry Automatic Rhythm and Style AnalyzeR, SLATE 2013 - Demonstration Track, Grenoble.

Delmonte R. & A.M. Prati. 2014. SPARSAR: An Expressive Poetry Reader, Proceedings of the Demonstrations at the 14th Conference of the EACL, Gotheborg, 73–76.

Delmonte R., 2015. Visualizing Poetry with SPARSAR - Poetic Maps from Poetic Content, Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature, Denver, Colorado, ACL, pp. 68–78.

Delmonte R., 2007. Computational Linguistic Text Processing – Logical Form, Semantic Interpretation, Discourse Relations and Question Answering, Nova Science Publishers, New York.

Delmonte R., 2009. Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution, Nova Science Publishers, New York.

Grosz, B. & Hirschberg, J. 1992. Some intonational characteristics of discourse structure. In *Proc. Int. Conf. Spoken Language Processing, Banff , Canada, 1986*, vol. 1, pp. 429-432.

Hamza W., Bakis, R., Eide, E.M., Picheny, M. A., & Pitrelli, J. F. (2004), The IBM Expressive Speech Synthesis System. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Jeju, South Korea, October, 2004.

Hirschberg, J. 1993. Pitch accent in context: predicting intonational prominence from text. Artificial Intell. 63, 305-340.

Huckvale M. 2002. Speech synthesis, speech simulation and speech science. In *Proceedings of the International Conference on Speech and Language Processing 2002*, pp. 1261– 1264.

Lakshmi Saheer, Blaise Potard, 2013. Understanding Factors in Emotion Perception, At 8° Speech Synthesis Workshop.

Lieske, C., J. Bos, M. Emele, B. Gamback, and C.J. Rupp 1997. Giving Prosody a Meaning, In Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech'97), Rhodes, Greece, 1431-1434.

Murray, I. R., & Arnott, J. L., 1993. Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustic Society of America, 93(2), pp. 1097–1108.

Nakatani, C. 1998. Constituent-based accent prediction. In *Proc. COLING-ACL '98, Montreal, Canada*, pp. 939-945.

Pierrehumbert, J. and J. Hirschberg (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack (Eds.), *Intentions in Communication*, pp. 271–311. Cambridge, Mass.: MIT Press.

Polzehl, T., S. Möller, and F. Metze, 2011. "Modeling speaker personality using voice," in Proc. INTERSPEECH, ISCA.

Prevost, S. 1995 A semantics of contrast and information structure for specifying intonation in spoken language generation. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA.

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

Shaikh, M. A. M., Prendinger, H., and Ishizuka, M., 2008. Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. Applied Artificial Intelligence, vol.22, issue 6, pp.558-601, Taylor & Francis.

Shimei Pan and Kathleen McKeown. 1997. Integrating language generation with speech synthesis in a Concept-to-Speech system. In *Proc. of ACL/EACL'97 Concept to Speech Workshop*, Madrid, Spain.

Shimei Pan and Kathleen McKeown. 1998. Learning intonation rules for concept to speech generation. In *Proc. of COLING/ACL'98*, Montreal, Canada.

Shimei Pan and Kathleen McKeown. 1999. Word informativeness and automatic pitch accent modeling. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., And Hirschberg, J. 1992. ToBI: A standard scheme for labelling prosody. In Proceedings of the International Conference on Speech and Language Processing 1992.

Steidl, S., T. Polzehl, T. Bunnell, Y. Dou, P. Muthukumar, D. Perry, K. Prahallad, C. Vaughn, A. Black, and F. Metze, 2012. Emotion identification for evaluation of synthesized emotional speech, in Proc. Speech Prosody.

Wang, M. & Hirschberg, J. 1992. Automatic classification of intonational phrase boundaries. Comp. Speech Lang. 6, 175-196..

Zovato, E., Francesca Tini-Brunozzi and Morena Danieli, 2008. Interplay between pragmatic and acoustic level to embody expressive cues in a Text to Speech system, in AISB Proceedings - Affective Language in Human and Machine, vol.2, 88-91.