An Annotation Scheme for Concept-to-Speech Synthesis

Janet Hitzeman[a,b], *Alan W. Black*[a], *Chris Mellish*[c], *Jon Oberlander*[b], *Massimo Poesio*[b] *and Paul Taylor*[a]

a. Centre for Speech Technology Research
b. Human Communication Research Centre
c. Department of Artificial Intelligence
2, Buccleuch Place
University of Edinburgh
Edinburgh, Scotland EH8 9LW

J.Hitzeman@ed.ac.uk

Word Count: 3504
Under consideration for other conferences (specify)? ACL

Abstract

The SOLE concept-to-speech system uses linguistic information provided by an NLG component to improve the intonation of synthetic speech. As the text is generated, the system automatically annotates the text with linguistic information using a set of XML tags which we have developed for this purpose. The annotation is then used by the synthesis component in producing the intonation. We describe the annotation system and discuss our choice of linguistic constructs to annotate.

An Annotation Scheme for Concept-to-Speech Synthesis

1 Introduction

The goal of the SOLE project is to make use of high-level linguistic information to improve the quality of the intonation of synthetic speech. SOLE's natural language generation (NLG) component automatically annotates the text it generates with high-level linguistic information using an XML-based annotation scheme, and it is this annotation which serves as the interface between the NLG component and the speech synthesis component of the system. In this paper we discuss our hypotheses concerning the correlation between the intonation and the linguistic constructs we chose to annotate and our motivation for choosing to use linguistic tags rather than annotating the text with the prosodic type and position of each accent.

2 The SOLE system

The SOLE concept-to-speech system is designed to work as a portable museum guide: visitors to a museum carry a portable device which detects what exhibits they are looking at and gives spoken explanation. SOLE generates its descriptions from a database of the museum exhibits' properties. As it keeps a record of which exhibits have already been visited, it is able to generate descriptions of new exhibits with reference to previous ones. This gives rise to a large number of linguistic phenomena such as various types of anaphoric reference (e.g., pronouns, definite descriptions, bridging references) and rhetorical relations (e.g., contrasting two exhibits or amplifying a particular property of an exhibit).

SOLE consists of an NLG component, a speech synthesis component and a set of XML tags which serves as an interface between the two systems. The NLG component of SOLE was developed for the ILEX project (Hitzeman et al., 1997), and currently it is used for describing exhibits in the Royal Museum of Scotland's Jewellery Gallery. The text-to-speech component is the Festival system.[1] The NLG component generates lexical, structural, semantic and discourse-level information concerning the text it generates, and automatically annotates the text with this linguistic information using the set of XML tags. The speech synthesis component makes use of these tags when determining what the intonation of an utterance should be. The current version of SOLE predicts the position of accents in the intonation contour. In the second phase of the project we will annotate the corpus with Tilt parameters (accent duration, amplitude, peak position, etc.) (Taylor, 1998) and we will also predict these values, thereby predicting not only the existence of an accent but also the size and shape of the accent.

The intonation component of Festival (Dusterhoff and Black, 1997) works by using a decision tree to analyse a set of features associated with a syllable, and to decide if a pitch accent should be assigned at that point. Typical features indicate whether a syllable has lexical stress, or part of speech of the word containing this syllable. In SOLE, we now have access to the higher-level linguistic information, and this greatly enriches the feature set that the decision

[1] http://www.cstr.ed.ac.uk/projects/festival.html

tree uses, e.g., we now have features indicating whether a syllable is contained in a noun phrase of type “anaphor” or in a rhetorical structure of type “contrast”.

In order to train the decision tree to use higher-level linguistic information in determining pitch accent placement, we needed a corpus consisting of the types of descriptive texts that the ILEX NLG system produces. At the time the SOLE project began, however, ILEX was in an early stage of development, so, rather than using ILEX output for our corpus, we gathered a corpus of texts of the sort that ILEX would be able to produce in its later stages. We annotated this corpus by hand with linguistic information, which involved deciding on an initial set of linguistic constructs that influence prosody and that can be produced by ILEX, and developing a set of XML tags to describe these constructs. We then recorded three speakers reading these texts, and human labellers marked accents on the speech by looking at the intonation contours. Given the annotated text and the accent annotation on the speech, we were able to extract the linguistic information on a per-syllable basis and use it as a set of features to train the decision tree to predict accent position.

3 The design of the annotation scheme

3.1 Linguistic tags vs. prosodic tags

Instead of using a set of linguistic tags, many concept-to-speech systems use tags that indicate the prosodic type of an accent directly (e.g., (Hiyakumoto et al., 1997)). For example, the annotation used by TrueTalk (Ent, 1995) can be used to specify ToBI type accents and speech rate, as illustrated in (1), where “H*1” is a pitch accent of size 1, and “r” specifies speech rate:

(1) `\!*H*1 Caution: \!pH*1 \!r1.05 \!*H*1 check for \!*H*1 hemoptysis`

The TraumaTalk concept-to-speech synthesis system (Bierner, 1998) uses a set of rules to map from its generated texts to this notation. However, while this prosodic annotation is interpretable by TrueTalk, it would have to be translated into some other prosodic annotation in order to be interpreted by other synthesisers, such as Festival; This mapping from one prosodic system to another is redundant. One concern of the SOLE project is that the system be synthesiser-independent, and therefore the most efficient choice of an annotation scheme for the output of the NLG component is a set of tags that simply represent the linguistic information produced by that component. This annotation requires only minor augmentation of the NLG system, compared with integrating the NLG system with a set of rules for prosodic markup. Given the linguistically annotated text, only one step of mapping to prosodic markup should be done, where the prosodic annotation is dependent upon the synthesiser.

Also, linguistic annotation allows for the mapping to prosodic markup to be done not only by a ruleset, as in TraumaTalk, but also via statistical methods. The rules are written according to a particular set of hypotheses, and they force the system to produce intonation according to those hypotheses. A statistical method, on the other hand, allows for the possibility of unexpected interaction between linguistic constructs and contributions from unexpected linguistic information. For example, our system finds that NPs expressing new information tend to be accented at the beginning of the phrase, which contradicts the standard claim that when a phrase is accented it’s accented at the end (Chomsky and Halle, 1968; Jackendoff, 1972).

There are also schemes for annotating linguistic information for concept-to-speech synthesis, such as (Pan and McKeown, 1997), but these schemes typically annotate syntax and possibly semantics, while our goal has been to annotate a much wider variety of linguistic information, as described in Section 3.3, below.

3.2 A note on reliability

Typically when developing an annotation scheme it is important to perform reliability studies to ensure that two human annotaters can reliably arrive at the same annotated form of a particular text (Bakeman and Gottman, 1997). However, for our purposes such studies are not crucial because, although we're currently using a hand-annotated training corpus, ultimately we will be using a corpus generated by our NLG component; Because this component generates the linguistic information as well as the text, it can annotate a text reliably 100% of the time.

3.3 Choice of linguistic constructs to annotate

Our first criterion for choosing linguistic constructs to annotate was the existence of some evidence that they contributed in interesting ways to intonation, and for this we consulted the psycholinguistic literature. Our second criterion was that these constructs had to be produced by our NLG system. Because our goal is not only to rely on existing hypotheses but also to explore new ones, and because the annotation is done automatically and is therefore cost-free, when we found a claim that a certain type of linguistic construct has an effect on intonation, we chose to annotate as many subtypes of that linguistic construct as our NLG system generates. For example, given a hypothesis that rhetorical structures of type "contrast" contribute to intonation, we chose to annotate every type of rhetorical structure that our NLG system produces.

We chose to annotate rhetorical structure, NP type (syntactic, semantic and reference type), topic/comment structure, syntactic structure, features of the text such as paragraph boundaries, punctuation, and parentheticals in this phase of the project. In the full paper we will discuss our hypotheses concerning the contributions that each of these linguistic components and constructs make to intonation; Here we will confine the discussion to rhetorical structure and NP type.

3.3.1 Rhetorical relations.

Rhetorical relations are discourse-level semantic relationships between segments of text. Some rhetorical relations, such as **contrast** and **list**, clearly have a corresponding intonational pattern; The effect that **contrast** has on intonation is commonly mentioned in the literature, e.g., (Chafe, 1974; Prevost and Steedman, 1994). With other types of rhetorical relations, such as **definition** and **exemplification**, the effect on intonation is not as obvious. Examples of the types of rhetorical relations we chose to annotate are below:

(2) **list, disjunction:** *[[Purple], [white] and [green]] are the colours of the suffragette movement.*

(3) **concession:** *[[This item is from the same period,] [but it doesn't have the same quality of workmanship.]]*

(4) **question:** *Was this item a tool?*

(5) **amplification:** *[This brooch is another example of figurative jewelry.] [In fact, the jeweller who made this piece made nothing but figurative jewelry for a number of years.]*
(6) **similarity:** *[Like the necklace designed by Flockinger,] [this item is in the Organic style.]*
(7) **contrast:** *[While the previous item was made to represent a Chinese god,] [this item was made to represent a Chinese demon.]*
(8) **definition:** *...the third is [a dress clip,] [which was used to fasten the straps of a dress at the neckline.]*
(9) **namely:** *...but from [an earlier period;] [around 1920.]*
(10) **exemplification:** *[The influence of textiles can be seen in a number of other pieces in the gallery,] [such as the Stick-On Butterfly designed by David Watkin and the tinted perspex earrings designed by Paula Dennet.]*

Each rhetorical structure can contain one or more **rhet-emph** tags, which mark the phrases within the text that express the properties or objects being compared, contrasted, listed, etc. We predicted that because these properties/objects have some kind of semantic emphasis that they would also be accented, and our results showed this to be the case. The following contrastive rhetorical structure illustrates our XML-based annotation:

(11) **<rhet-elem type="contrast">**
  <nucleus> *The*
    <rhet-emph type="object"> god **</rhet-emph>**
    was
    <rhet-emph type="property"> *gilded* **</rhet-emph>**;
  </nucleus>
  <nucleus> *the*
    <rhet-emph type="object"> *demon* **</rhet-emph>**
    was
    <rhet-emph type="property">
      stained in black ink and polished to a high sheen
    </rhet-emph>.
  </nucleus>
</rhet-elem>

Because we are only concerned with predicting accent placement in the first phase of the SOLE project, the rhetorical emphasis (**rhet-emph**) is the only relevant annotation; the rhetorical structure type, rhetorical emphasis type and the nuclei and satellites will be important when predicting tune in the next phase of the project.

3.3.2 Noun phrases.

It is well known that old information tends to be deaccented and new information tends to be accented (Chafe, 1974; Crystal, 1975). The first time an object is mentioned in a text it is part of the new information in that text, and all subsequent references to that object are considered references to old information, as illustrated in (12):

(12) *It was worn mainly by teenagers, to show that they were Beatles fans, or perhaps to show which of the Beatles they liked best.*

The first time the NP *Beatles* is mentioned in the text, it is new and likely to be accented; the subsequent reference to the Beatles refers to old information, and is unlikely to be accented.

Making use of old and new information is becoming more common in concept-to-speech systems (e.g., (Prevost, 1995; Hiyakumoto et al., 1997; Nakatani, 1997)). We chose a more complex annotation scheme for NPs, assigning them a reference type, a syntactic type and an optional semantic type. These attributes are described below.

3.4 Reference type

As NPs, all of the items tagged as **np-elem**s have the potential

- to refer to something (an entity, event, description, etc.) previously mentioned (**anaphor**),
- to act as a predicate, saying what the entity under discussion is (**predicative**),
- to refer to entities closely related to entities previously mentioned (**bridge**, **explicit-bridge**),
- to introduce a new entity into the discourse (**first-mention**) or
- to refer to something that hasn't been mentioned but that is known to the reader/hearer to be in the situation under discussion (**situation**).

Examples of the reference types are given below:

(13) · first-mention *(worn mainly by [teenagers])*
· anaphor *(it)*
· bridge *(although [the interior] is smooth)*
· explicit-bridge *(its hollow terminals of sheet gold)*
· predicative *(This item is [a bronze ritual food vessel])*
· situation *(the 1960s, sits on [your neck])*
· kind-of *(a dress-fastener.... [The dress-fasteners]....)*
· instance-of *(hoards.... [The Fishpool Hoard]....)*

It is generally accepted that **anaphor**s, which express old information (in the sense that they've been mentioned before), tend to be deaccented, and that **first-mention**s, which express new information, tend to be accented. Bridging references can be thought of as new information because they refer to entities not previously mentioned, but they can also be thought of as old information because they are related to an entity that has been mentioned, so it is not clear whether we ought to hypothesise that they are accented or not. It's also unclear whether references to the current situation will be accented or not, because they are both new information (since they haven't been mentioned previously) and old information (since both the writer and reader are aware of their presence in the current situation).

3.5 Syntactic type

It is likely that **syn-type** will have little or no effect on intonation. The claims in the literature are that whether an NP is accented or not depends on whether it expresses old or new information, which predicts reference type should be the strongest indicator of a particular NP's accent. However, given that our NLG system can easily annotate each NP with its syntactic type, it makes sense to test for correlations between syntactic type and intonation, even if all we do is confirm that there are none.

(14) · indefinite-NP *(a jewel)*
· definite-NP *(the jewel)*
· possessive-NP *(King's brooch)*
· conj *(solidarity or affiliation)*
· gerund *(knitting)*
· bare-plural *(jewels)*
· deictic-NP *(this jewel)*
· deictic-pro *(this)*
· reflexive-pro *(itself)*
· possessive-pro *(their)*
· personal-pro *(she)*
· quantified-NP *([most costume jewellery])*
· N-bar *(most [costume jewellery])*
· bare-NP *(the [suffragette] movement)*
· weak-indefinite *(a piece of jewellery)*
· bare-singular *(jewellery, silver)*
· mass-indefinite *(wire)*

3.6 Semantic type

(15) · PN *(John, the Middle Ages)*
· kind *(jewellery, a type of brooch)*

It may be the case that proper names in English are spoken with a certain rhythm and a certain duration. For example, we assign the name *John Smith* two beats and a particular duration, and speed up in order to give *Miriam Hardcastle* the same duration and beats. It may be the case that names of people have different intonation than other types of proper names such as *the Middle Ages*, so we should distinguish cases in which the **PN** has syntactic type equal to **bare-singular** from other cases of **PN**s.

An example of the annotation is in (16):

(16) was
<**np-elem ref-type="predicative" syn-type="indefinite-NP"**>
an
<**np-elem ref-type="first-mention"**
syn-type="N-modifier" sem-type="PN">
Edinburgh
</**np-elem**>
jeweller
</**np-elem**>

4 Modularity of the annotation scheme

We chose an annotation scheme that would allow the interaction between linguistic modules to be clearly seen. In our XML DTD, we have one element for each linguistic module, i.e., one each for syntax, semantics, rhetorical structure, topic/comment structure, etc. Our statistical method shows which of these modules interact by finding elements that combine to form good predictors of accenting. Again, the statistical method may give surprising results; For example, we hypothesise that topics tend to be deaccented (Chafe, 1974) but that they will be accented if they are in a contrastive emphasis. The statistical method may also find that they are accented when contained in a particular syntactic position or semantic construct, or it may predict that only topics of a certain np-type are deaccented.

5 Results and summary

We have described an annotation scheme for concept-to-speech synthesis and have argued that using linguistic tags rather than prosodic tags is more effficient in a synthesiser-independent concept-to-speech system. This type of tag also allows for the use of a statistical method in determining the mapping from linguistic representation to intonation, which makes it possible for the system to discover generalisations not found in the literature.

Our results show that the addition of NP and rhetorical structure annotation alone reduces the error in accent prediction by 15.5%. The strongest results are that first-mention NPs and **rhet-emph** phrases are good indicators of accenting, and one unexpected result for our corpus is that first-mention NPs tend to be accented at the beginning of the phrase rather than the end.

References

R. Bakeman and J. M. Gottman. 1997. *Observing Interaction: An Introduction to Sequential Analysis.* Cambridge University Press, Cambridge, second edition.

Gann Bierner. 1998. Traumatalk: Content-to-speech generation for decision support at point of care. In *A Paradigm Shift in Health Care Information Systems.* AMIA, November.

Wallace L. Chafe. 1974. Language and consciousness. *Language*, 50:111–133.

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

David Crystal. 1975. Prosodic features and linguistic theory. In *The English tone of voice: Essays in intonation, prosody and paralanguage*, pages 1–46. Edward Arnold, London.

K. Dusterhoff and A. W. Black. 1997. Generating F0 contours for speech synthesis using the Tilt intonation theory. In *Proceedings of the ESCA Workshop on Intonation*, Athens, Greece.

Entropic Research Laboratory, 1995. *TrueTalk Reference Manual*.

Janet Hitzeman, Chris Mellish, and Jon Oberlander. 1997. Dynamic generation of museum web pages: The intelligent labelling explorer. *Archives and Museum Informatics*, 11:107–115. Also presented at the Museums and the Web Conference, Los Angeles, March 1997.

Laurie Hiyakumoto, Scott Prevost, and Justine Cassell. 1997. Semantic and discourse information for text-to-speech intonation. In Kai Alter, Hannes Pirker, and Wolfgang Finkler, editors, *Concept to Speech Generation Systems*, pages 47–56, Madrid, Spain, July. Association for Computational Linguistics.

R. Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA.

Christine H. Nakatani. 1997. Computing prosody. In *Integrating Prosodic and Discourse Modelling*. Springer-Verlag.

Shimei Pan and Kathleen R. McKeown. 1997. Integrating language generation with speech synthesis in a concept to speech system. In Kai Alter, Hannes Pirker, and Wolfgang Finkler, editors, *Concept to Speech Generation Systems*, pages 23–28, Madrid, Spain, July. Association for Computational Linguistics.

Scott Prevost and Mark Steedman. 1994. Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153.

Scott Allan Prevost. 1995. *Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. Ph.D. thesis, University of Pennsylvania.

Paul Taylor. 1998. The Tilt intonation model. In *International Conference on Spoken Language Processing (ICSLP)*, December. Paper number 827.