ON THE USE OF AUTOMATICALLY GENERATED DISCOURSE-LEVEL INFORMATION IN A CONCEPT-TO-SPEECH SYNTHESIS SYSTEM

Janet Hitzeman[ab] *Alan W. Black*[a] *Paul Taylor*[a] *Chris Mellish*[c] *Jon Oberlander*[b]

a. Centre for Speech Technology Research
b. Human Communication Research Centre
c. Department of Artificial Intelligence
University of Edinburgh
Edinburgh EH1 1HN, GB
http://www.cstr.ed.ac.uk/projects/sole.html email: J.Hitzeman@ed.ac.uk

ABSTRACT

This paper describes the latest version of the SOLE concept-to-speech system, which uses linguistic information provided by a natural language generation system to improve the prosody of synthetic speech. We discuss the types of linguistic information that prove most useful and the implications for text-to-speech systems.

1. INTRODUCTION

The purpose of the SOLE project is to make use of automatically-generated, high-level linguistic information to improve the quality of the intonation of synthetic speech. After choosing an initial set of linguistic constructs thought to have some influence on prosody, we developed an SGML-based mark-up language to serve as a general interface between NLG and speech synthesis systems, and trained our synthesis system to recognise correlations between the mark-up and intonational contours so that it can make use of this mark-up when synthesising. As a result, many of the errors that the synthesiser makes with regard to knowing when to accent or deaccent a word are absent in the SOLE output. This paper reports on the current results and discusses the implications for text-to-speech systems in cases where it is realistic to use statistical methods for exploiting certain types of high-level linguistic information.

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 what 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 discourse-level 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).

The NLG component of SOLE was developed for the ILEX project [5], and currently it is used for describing exhibits in the Royal Museum of Scotland's National Jewellery Gallery. The text-to-speech component is the Festival system. [1]

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

The intonation component of Festival [4] 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 used include lexical stress and position in phrase etc. In SOLE, we now have access to the discourse-level information, and this greatly enriches the feature set that the decision tree uses.

3. METHOD

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 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. Our corpus consists of 43 short descriptive texts, which gives us 35 minutes of speech and a total of 6331 syllables, 863 of which we set aside for testing. We annotated this corpus 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 SGML 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 F0 contours.

Given the tagged text, 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. The SOLE NLG component (i.e., the augmented ILEX system) automatically produces tagged text, which the trained decision tree then uses in determining accent placement. In the second phase of the project we will annotate the corpus with Tilt parameters [11] (accent duration, amplitude, peak position, etc.) and we will also predict these values.

Of the phenomena we chose to annotate in the first phase of the project, noun phrases (with their syntactic, semantic and reference type) and rhetorical structure gave the most significant results for accent placement, so we will restrict our discussion here to these constructs.

3.1. Linguistic annotation

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; with others, such as **definition** and **exemplification**, the effect on intonation is not as obvious. Examples of a few of the types of rhetorical relations we chose to annotate are below:

(1) **List:** *[Purple, white and green] were the colours of the suffragette movement.*
(2) **Similarity:** *[[Like the necklace designed by Flockinger,] [this item is in the Organic style.]]*
(3) **Concession:** *[[This item is from the same period,] [but it doesn't have the same quality of workmanship.]]*

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. The following contrastive rhetorical structure illustrates our SGML-based annotation:

(4) <**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.

Noun phrases. It is well known that old information tends to be deaccented and new information tends to be accented [1, 3]. 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 5:

(5) *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., [9, 6, 8]). We chose a more complex annotation scheme for NPs, assigning them a reference type, a syntactic type and an optional semantic type.

In addition to annotating NPs as **anaphor**s (old information) and **first-mention**s (new information), we used a third reference type, **predicative**, illustrated in 6:

(6) *This item is [a brooch].*

A predicative NP is one that generally occurs as the object of *to be*, giving a description of the subject.

Among the syntactic types we assigned to NPs are the following:

- **definite NP**: Any NP using the definite determiner (*the*).
  - *the brooch, the north-west portion of the coastline of the Firth of Forth*
- **bare-singular**: A singular NP without a determiner.
  - *jewellery, 1920, purple, solidarity*
- **N modifier**: A noun that modifies the head noun in a noun-noun compound.
  - *[costume] jewellery, a [dress] clip, an [Edinburgh] jeweller*

The semantic types we chose to annotate are below:

- **proper name**: E.g., *Jesse M. King, Scotland, the Middle Ages*
- **kind**: An NP that describes a kind of object rather than an instance of an object.
  - *jewellery, people, the mass-produced variety of jewelry which was popular during the 1930s*

An example of the annotation is in 7 (Note that the term **anaphora-elem** could be replaced with **noun-phrase**):

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

4. RESULTS

Table 1 gives a comparison of the number of errors made by the TTS system using the original set of features with the number of errors made when the SOLE linguistic features were added to the set. Overall, the addition of linguistic features reduces the error in accent prediction by 15.5%. The features in Table 1 show the largest contribution to the error reduction.

The first two features in Table 1 are purely syntactic indicators of whether a syllable is in an NP or an embedded NP. This simple classification isn't very useful, as shown by the small reduction in error

Syllable feature	Total occurrences	TTS errors	TTS + SOLE errors	% error reduction
any syl in an NP	601	88	85	3.4
any syl in embedded NP	213	33	29	12.1
any syl in an anaphor	135	22	3	86.4
last stressed syl in anaphor	41	12	9	25.0
any syl in a first-mention	276	35	6	82.9
last stressed syl in first-mention	54	11	6	45.5
any syl in a predicative NP	69	15	5	66.7
any syl in a definite NP	153	18	1	94.4
any syl in a bare-singular	114	24	16	33.3
any syl in an N-modifier	12	3	0	100.0
any syl in a deictic NP	52	5	3	40.0
any syl in a proper name	114	21	0	100.0
first stressed syl in a proper name	37	10	0	100.0
any syl in a kind	77	7	0	100.0
any syl in **rhet-emph**	678	104	94	9.6
last stressed syl in **rhet-emph**	46	12	5	58.3

Table 1: A comparison of the TTS system with the SOLE system

for both features. However, the next four features, which include information concerning the reference type of the NP, show large reductions in error. As expected, anaphors tend to be deaccented and first-mentions to be accented. What is unexpected is that our results contradict the general claim in the literature that when a phrase is accented the accent is placed at the end of that phrase [2, 7]: in our data, the feature indicating that a syllable is *any* (stressed) syllable in a first-mention proves more useful than the feature indicating that a syllable is the last stressed syllable in a first-mention.[2]

The next eight features also give unexpected results; Predicative NPs, definite NPs, bare singular NPs, N-modifiers, deictic NPs, proper names and kinds are not typically spoken of as indicators of accenting. With N-modifiers and deictic NPs, admittedly, the numbers are small and it is therefore difficult to make a strong argument that they will be reliable indicators of accenting in another corpus. However, for the other features the reduction in error is large. Predicative NPs express new information, which explains the observation that they tend to be accented; the discovery that an NP following the verb "is" or "are" is likely to be accented has strong implications for TTS systems with shallow statistical parsing mechanisms. In contrast, definite NPs generally express old information because they refer to objects previously mentioned in the text;[3] again, the implication is that a TTS system could predict deaccenting on an NP beginning with *the*. Proper names, bare singular NPs and kinds can either express old or new information, so it is surprising that they serve as indicators of accenting. Also surprising is that it is useful to know whether a syllable is the *first* lexically stressed syllable in a proper name.

[2]Our decision tree always predicts that a syllable without lexical stress is deaccented, and the linguistic features are therefore only used in deciding whether a stressed syllable is accented.

[3]There are exceptions, such as definites that refer to objects in the common ground, as in "[The weather] is lousy today", and bridging references, which are definites that refer to an object closely related to a previously mentioned object, as in "There is a house on the hill. [The door] is green." There is no consensus on whether these constructs describe old or new information.

The last two features concern rhetorical structure. Both using a feature indicating whether a syllable is inside a **rhet-emph** tag and using a feature indicating whether a particular syllable is the last lexically stressed syllable in that tag gave a reduction in error. The reduction in error is greater with the latter feature, indicating that in rhetorical emphases the accent tends to be placed at the rightmost portion of the emphasised phrase.

The type of rhetorical structure had negligible effect on accent placement, although, unsurprisingly, the phrases with rhetorical emphasis had a tendency to be accented. We predict that because rhetorical structures have tunes which are dependent on their type, their type will be important in the next phase of the project, when we train the system to recognise other features of an accent, such as duration and amplitude.

5. DISCUSSION

There are three central conclusions to be drawn from our results:

1. In the domain of descriptive texts, certain types of high-level linguistic information are useful in determining accent placement, and therefore coupling a natural language generation system with a speech synthesis system is a good idea;
2. Surprisingly, kinds, bare-singular NPs and proper names are good predictors of accents; and
3. A TTS system with a statistical parsing mechanism would benefit from singling out predicative NPs, definite NPs and bare-singular NPs (both proper names and kinds) because they are easily recognised via statistical methods and are good predictors of accent placement.

Another goal of the SOLE project is to formalise the provision of discourse-level information as a set of SGML tags as part of the standardisation efforts of the SABLE consortium [10]. The intention here is to design a powerful interface language between language generation and speech synthesis systems, so that the synthesis

systems can produce high quality speech in a variety of applications and domains.

6. Acknowledgements

SOLE is funded by the EPSRC, grant reference GR/L50341.

7. REFERENCES

1. Wallace L. Chafe. Language and consciousness. *Language*, 50:111–133, 1974.
2. N. Chomsky and M. Halle. *The Sound Pattern of English*. Harper and Row, New York, 1968.
3. David Crystal. *Prosodic features and linguistic theory*, pages 1–46. The English tone of voice: Essays in intonation, prosody and paralanguage. Edward Arnold, London, 1975.
4. K. Dusterhoff and A. W. Black. Generating F0 contours for speech synthesis using the Tilt intonation theory. In *Proceedings of the ESCA Workshop on Intonation*, Athens, Greece, 1997.
5. Janet Hitzeman, Chris Mellish, and Jon Oberlander. Dynamic generation of museum web pages: The intelligent labelling explorer. *Archives and Museum Informatics*, 11:107–115, 1997. Also presented at the Museums and the Web Conference, Los Angeles, March 1997.
6. Laurie Hiyakumoto, Scott Prevost, and Justine Cassell. 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 1997. Proceedings of a Workshop Sponsored by the Association for Computational Linguistics.
7. R. Jackendoff. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA, 1972.
8. Christine H. Nakatani. *Computing Prosody*, Integrating Prosodic and Discourse Modelling. Springer-Verlag, 1997.
9. Scott Allan Prevost. *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. PhD thesis, University of Pennsylvania, 1995.
10. R. Sproat, A. Hunt, M. Ostendorf, P. Taylor, A. W. Black, K. Lenzo, and M. Edginton. *SABLE: A standard for TTS markup*. ICSLP98, this volume.
11. P. Taylor. *The Tilt Intonation Model.* ICSLP98, this volume.