On the use of automatically generated discourse-level information in a concept-to-speech synthesis system

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney Convention Centre, Sydney, Australia, 30th November - 4th December 1998

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
ON THE USE OF AUTOMATICALLY GENERATED DISCOURSE-LEVEL INFORMATION IN A CONCEPT-TO-SPEECH SYNTHESIS SYSTEM

Janet Hitzeman\textsuperscript{a,b}  
Alan W. Black\textsuperscript{a}  
Paul Taylor\textsuperscript{a}  
Chris Mellish\textsuperscript{c}  
Jon Oberlander\textsuperscript{b}

\begin{itemize}
\item[a] Centre for Speech Technology Research
\item[b] Human Communication Research Centre
\item[c] Department of Artificial Intelligence
\end{itemize}
University of Edinburgh
Edinburgh EH1 1HN, GB

\url{http://www.cstr.ed.ac.uk/projects/sole.html}  
email: J.Hitzeman@ed.ac.uk

\section*{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.

\section{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.

\section{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. \footnote{\url{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.

\section{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 Tits 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.

\subsection{Linguistic annotation}

\textbf{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 god was gilded. </nucleus>
    <nucleus> the demon was stained in black ink and polished to a high sheen. </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 **anaphors** (old information) and **first-mentions** (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 turned into a functional element:
    an <anaphora-elem ref-type="predicative" syn-type="indefinite-NP">
        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...
Table 1: A comparison of the TTS system with the SOLE system

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

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