Prosody Transfer in a Spoken Dialogue System *

W. N. Campbell & Pao-Chen Hwang
ATR Interpreting Telecommunications Research Laboratories
Hikari-dai 2-2, Seika-cho, Kyoto 619-02 Japan.

Abstract

This paper describes experimental work in the design of a prosody transfer process for a speech translation system. We describe the architecture and discuss the relevance of prosodic information flow in the translation of sentences with different focus.

1 Introduction

As a preliminary step in the integration of speech recognition, translation, and synthesis components of a spoken dialogue translation system, we have implemented a prosody transfer module for the transmission of focus information between the spoken input and the generated utterance. The text-to-speech component in a dialogue system has more information about how the utterance should be produced than does a simple text-based synthesiser so we complement the language-translation module by using prominence information extracted from the original utterance as additional input to the speech synthesiser. We follow the translation tree to find equivalent words in the translated sentence in order to carry across focus information and generate appropriate prosody to signal the intended meaning in the translated output synthesis.

2 Information flow

The software consists of a series of pesudo-terminals running on the same or different machines to interface a stand-alone speech recogniser and TDMT [2], the ATR transfer-based translation software, with CHATR a multilingual speech synthesiser [3], in a socket-based server-client relationship.

Figure 1 shows taps from the acoustic parameters of the speech recogniser to extract F_0 and power information, and from the translator for text mappings. In the current implementation, prosodic information is not yet fed directly into the TDMT translation component. Figure 2 shows a detail of the prosody analyser, where three mappings take place.

*音声翻訳の韻律トランスファー by ニック キャンベル、黄宝堉。

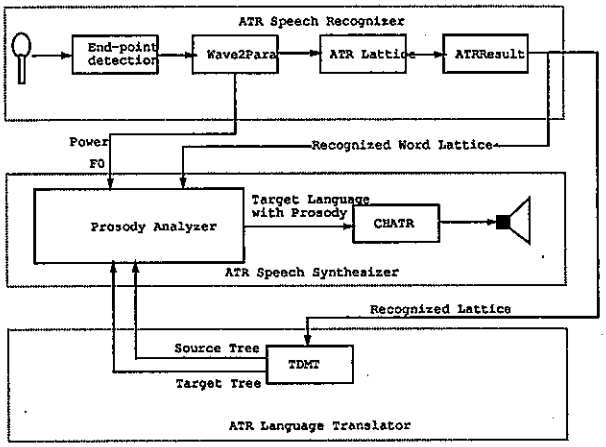

Fig. 1 Information flow for prosody transfer

3 Prosody transfer

As an example of the prosody transfer process, consider the utterance: "Could I have your phone number too?" (see Figure 3). Focus could fall naturally on words 2, 4, 5, 5&6, or 7, or in the marked case on any word. Focus is not determinable from the text. The scope of 'too' could be limited to 'phone-number' or could extend to include 'your' or 'I'. In speech, such differences are normally indicated by prosody.

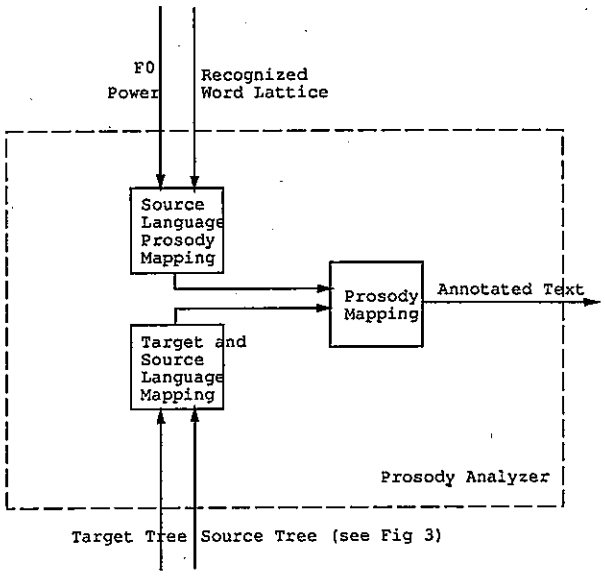

Fig. 2 Three transfer mappings

```
Input: "Could I have your phone number, too?"
                        ^^^^^^^^^^^^

=> TOP [(?X too) --- AS]
   |--?X [(could ?X <PRON-V> ?Y) --- MS]
      |--?X [i]
      |
      |--?Y [(?X <V-DET> ?Y) --- PN]
         |--?X [(have)]
         |
         |--?Y [(your ?X) --- DN]
            |--?X [(phone number)]

=> TOP [0.000 ((副詞 また) {読点_ 区切} !X)]
   |--!X [0.000 ({_ 副詞節} !Y (接続助詞 ても)
                    (形容詞 よい) (終助詞 か))]
      |--!Y [0.000 (!Y (格助詞 を) !X)]
         |--!Y [0.000 ((接頭辞 御) !X)]
         |  |--!X [STRING ((普通名詞 電話番号))]
         |
         |--!X [STRING ((本動詞 伺う))]

  (副詞 "また"){読点_ 区切}
  ({_ 副詞節}
   (接頭辞 "御")
   (普通名詞 "電話番号")
   (格助詞 "を")
   (本動詞 "伺う")
   (接続助詞 "ても")
   (形容詞 "よい")
   (終助詞 "か")))

=> また、御電話番号を伺ってもよいですか。
          ^^^^^^^
```

Fig. 3 Following the translation tree to identify the target nominal in the other language.

After translation by TDMT, the sentence becomes "また、御電話番号を伺ってもよいですか。" and in the default case a small focal accent is placed on "電話番号". Should focus have been detected on 'I' (me too) then a different translation would have been called for.

To determine which word or phrase is in focus, a prominence detector [1] is run for each word in the input using the acoustic information and times from the recogniser, and the scores are carried across the transfer tree to match equivalent words in the output.

The synthesiser has been programmed to modify the fundamental frequency and power of the generated speech according to the prominence score on each word.

4 Evaluation of the system

A test was carried out to confirm that native listeners were able to recognise the intended difference in the synthesised output: 10 sentences from a conference registration task were repeated three times each, with focal prominence on different words within the utterance each time, giving 3×10×3=90 utterances in all. A panel of 10 native listeners of Japanese were able to correctly detect focal prominence in 74% of the synthesised sentences. For the same sentences produced by a native speaker of Tokyo Japanese, correct recognition of intended focus was 84%. Since the chance score is 33%, we can conclude that it is clearly worthwhile to add focus information in this way to the synthesised output.

5 Future work

This work represents only a first step towards implementing prosody transfer in a translation system and is limited to the transmission of prominence in a utterance where more than one nominal group can be in focus. Future work must incorporate prosodic information directly into the linguistic component of the translation, both to feed higher-level intonation prediction within the synthesiser, and to trigger alternative translations when an ambiguity in e.g., scope or phrasing can be resolved by the extra information from the non-textual input.

Further interesting work remains to be done at a completely different level on the differences between languages in what they express by prosody. As a first step though, Japanese and English appear to have enough similarities to make continued experimentation worthwhile.

6 Conclusion

We have described the basic architecture for a prosody transfer component of a spoken-language translation system. Operating in parallel with the linguistic component it takes input directly from the speech recogniser, combining acoustic parameters time-synchronously with the text, and mapping from source to target language to indicate through the output synthesis which word or phrase carried focal prominence in the original speech. Tests with synthesised utterances confirmed that listeners were able to correctly identify the intended focus at rates significantly better than chance.

参考文献

[1] Ding W., & Campbell, N., "On the correlation of prominence and voice-source", this volume.

[2] Furuse, O, et.al. "Transfer driven machine translation utilizing empirical knowledge", Trans IPS Japan, 35 pp 414-425, 1994.

[3] Campbell, N., "High-definition speech synthesis", this volume.