A practical Message-to-Speech strategy for dialogue systems

P. Spyns (1), **F. Deprez** (1), **L. Van Tichelen** (1) and **B. Van Coile** (1,2)
(1) Lernout & Hauspie Speech Products, Sint Krispijnstraat 7, B-8900 Ieper, Belgium
tel.: 32-57-22.88.88, fax: 32-57-20.84.89
(2) E.L.I.S., University of Gent, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
tel.: 32-9-264.33.95, fax: 32-9-264.35.94
{Peter.Spyns,Filip.Deprez,Luc.VanTichelen,Bert.VanCoile}@lhs.be

Abstract

In this paper, we present a Message-to-Speech system for Natural Language Generation that is to be integrated in a dialogue system. As the system has to function in a very restrictive environment with respect to computational resources, a compromise between concept based and template based generation systems had to be found. Still, the approach aims at achieving linguistic flexibility for the utterances and attaining a natural sounding prosody.

1 Introduction

Many of the Natural Language Generation (NLG) systems that produce flexible output, i.e. sentences with variations on the syntactical and morphological levels, only aim at the production of written text and do not deal with spoken language. By doing so, the important topic of generation of natural prosody is not touched upon (see e.g. (Elhadad, 1992; Reiter et al., 1995; Dalianis, 1996; Somers et al., 1997)). On the other hand, message generating systems that provide speech of a natural quality (e.g. announcement systems, phone banking and voice mail applications) often combine fixed pieces of pre-recorded speech. These text and message generating systems are either resource intensive (powerful CPU, large storage and memory capacity, ...) or provide only limited flexibility, which seriously hampers their integration in a dialogue system.

The Message-to-Speech (MTS) system described below is specifically designed to function in an environment with seriously restrained computational resources where it is impossible to store large amounts of pre-recorded speech. In this context, Text-to-Speech (TTS) is an evident alternative. However, for dialogue systems using a predefined set of message types, the use of special purpose prosody models can lead to a prosodic quality that is superior to the one generated by TTS systems, which apply general purpose prosody models for unrestricted text (see also (Hovy, 1995, p.161)). Our prosody transplantation tool (see section 2) exploits this idea: for the fixed parts of a message it allows to overrule prosody generated by general models, as is done by TTS, with specific prosody copied from natural speech. Prosody by general model is only used for those parts of the message where flexibility is needed. The MTS system combines transplanted prosody with prosody by model in order to cope with partly variable messages while still preserving natural prosody (Van Coile et al., 1995).

Details on the MTS system will be provided in the third section. It consists of two components: the MTS generation and the MTS prosodic integration parts. The former module (see section 3.1) is template driven (canned "text" interspersed with slots). For a discussion of template driven systems see (van Deemter et al., 1994; van Deemter and Odijk, 1997; Reiter, 1995). The templates account for the flexibility, including the linguistic variation, of the messages. The latter module (see section 3.2) specifically takes care of assimilation and the prosodic integration of the slot values with the rest of the template. A discussion concludes this paper (see section4).

2 Prosody Transplantation

The idea behind Prosody Transplantation is that of copying intonation and duration values from a recorded *donor* message (human speech) to the phonetic transcription of the same message. The specific *Enriched Phonetic Transcription* (EPT) obtained in this manner can be fed to a TTS system whereby the normal linguistic and prosodic modules (based on general models) are by-passed (Phonetics-to-Speech — PTS). Only the segmental synthesis and the synthesiser modules are used.

An example of an EPT is provided by figure 1. The first value between square brackets is the phoneme duration (in ms), optionally followed by one or more intonation breakpoints. Each breakpoint consists of a location value (in ms) relative to the beginning of the phoneme, followed by a pitch

```
# T[104] æ[74(0,98)] N[47] k[107(10,81)] j[14(0,106)]
u[44] f[93(0,91)] o[47(0,102)] r[29] j[68(0,98)(30,90)]
o[50(0,96)] r[71] $[45(0,93)] -t[108] E[70(0,102)] n[68]
-S[96] $[56] n[106(30,83)(100,83)] #
```

Figure 1: textual representation of an EPT for the sentence "Thank you for your attention"

value (in ST/4; reference 50 Hz).

A major asset of Prosody Transplantation is the combination of natural sounding speech with a low bit rate for storage (less than 300 bit per second). In addition, only the prosody and not the timbre of the speaker is retained. New donor messages can be recorded by new speakers and seamlessly integrated in existing applications. Specific tools have been developed to speed up the prosody transplantation process (Van Coile et al., 1994). Although the EPTs as such do not support linguistic variation, the combination of PTS with a template driven system provides linguistic flexibility as well as natural prosody.

3 The Message-to-Speech System

In the following sections, more details will be provided about the combination of fixed and variable information (templates and arguments). Once the appropriate surface form is selected (see section 3.1), the resulting EPT template with its arguments (phonetically transcribed) is integrated on the prosodic level (see section 3.2). Finally, the integrated EPT is fed into the TTS synthesis module (PTS).

3.1 MTS Generation Module

3.1.1 Basic Concepts

A *message* represents a complete sentence and is composed of one or more building blocks or *message units* (MU), which constitute the input of the MTS system. All MUs are prosodic units that cannot be combined in an arbitrary way to form messages: syntax specifies how to combine different MUs units into a message. The flexibility of a MU is guaranteed by the presence of slots. By providing different arguments for a slot, several variants can be derived from the same MU at run-time. An entire message can thus be parameterised.

Subsequently, the MUs are mapped into one or more *carriers*. A carrier is a template containing the enriched phonetic transcription of canned text, transplanted from an appropriate donor message (see above), together with the prosodic information for the free slot parts (see below).

MU arguments are not necessarily passed on to a carrier slot in a straightforward way: the argument can be deleted, adapted, or swapped. Examples of MUs and carriers are given in figure 2 [1].

[1] Although some examples show an orthographic representation, it must be stressed that a carrier is a very concise representation of a piece of recorded speech without segmental voice-specific features. Each phoneme also has duration and intonation characteristics (see figure 1).

3.1.2 Basic Algorithm

The MTS generation part basically tries to procure a method that ensures the variability of a piece of information and takes the related linguistic variations into account (selection of the correct variant). The transformation of a MU into one or more carriers is guided by a two-fold mechanism:

- argument dependent carrier selection: the carrier is selected in function of (a characteristic of) an argument. E.g. /a/ car_ vs. /two/ cars (singular vs. plural templates). In order to select the appropriate carrier, morpho-syntactic information about the argument must be available (in a dictionary) .
- carrier dependent argument realisation: the argument is realised in a different way in function of the selected carrier. E.g. /a/ car vs. /an/ automobile (vocalic onset or not for singular noun). For the argument to be realised correctly, linguistic constraints on the slot must be taken into account.

The arguments to be filled in a slot are phonetic transcriptions provided by a dictionary or a grapheme to phomene (G2P) conversion module. E.g., the dictionary entry for the determiner is *an;ON=VO | a,NB=SG*: "a" is the default; "an" is used before nouns with a vocalic onset and both forms are singular. It will be clear that the prosody of a carrier (EPT with slots), although better than plain TTS, risks to be slightly inferior to that of an entire EPT (no slot). Therefore, a good and practical compromise has to be found for the trade-off between storage space on the one hand and flexibility and prosodic quality on the other .

An example (see figure 3) gives an idea of how the system works. The transformation of MU 0001 into carrier 3551 is straightforward (no specific condition). Depending on the value of the argument

	ID	ARG	Comment
MU	0001		Welcome to ...
carrier	3551		
MU	0002		in /X/ mile
carrier	3561,%1	SG	in /a/ mile
carrier	3562,%1	*	in /Y/ miles
MU	0003		turn /LeftRight/
carrier	3571	left	turn left
carrier	3572	right	turn right

Figure 3: example conversion table of MUs into carriers

	type of mapping	message unit (MU)	carrier (orthographic representation)
1	1 MU to 1 carrier	welcome to the navigation system	Welcome to the navigation system.
2	1 MU to 2 carriers	in */num/* mile */s/*	In /a/ mile_, In */num* $\neq$ *1/* miles,
3	1 MU with 1 slot to 2 carriers without slot	turn */LeftRight/*	Turn right. Turn left.

Figure 2: example of mapping of message units (MU) to carriers

(ARG), MU 0002 is mapped onto carrier 3561 or carrier 3562.

This is an example of argument dependent carrier selection. Subsequently, if alternative surface forms co-exist, the restriction on the slot (see figure 4) is compared with the characteristics of its argument. As "an" is associated with "ON=VO" (vocalic onset), the default case "a" is selected (= carrier dependent argument realisation).

CARRIER : 3562 In /Distance/ miles
[952(952,101)] ?[18] I[66] n[92(4,98)]
/Distance: ... **ON=CO** .../
m[138(10,103)(70,96)] Y[224(2,93)(132,92)]
l[173(58,82)] z[352] #[411(231,82)]

Figure 4: example of a carrier (with a morphological restriction on the slot: onset is not vocalic [ON=CO])

3.2 MTS Prosodic Integration Module

The purpose of the prosodic integration module is to calculate appropriate prosody for all *arguments* that are to be filled out in a carrier. In a first step a duration is calculated for each of the phonemes in the argument (see section 3.2.1). In a second step, an appropriate intonation contour is calculated (see section 3.2.2).

3.2.1 Duration module

The input of the duration module is a phonetic transcription in which primary and secondary stress, provided by the dictionary or G2P module, are indicated. The duration module has access to one or more duration models in order to produce a phonetic transcription that is enriched with a duration value for each phoneme.

A duration model is a rule-based system calculating durations, taking into account parameters such as lexical stress, position of phonemes (word initial, word medial, word final, sentence final), length of the argument, phonetic context of phonemes (left/right neighbour, consonant cluster, open/closed syllable) etc. As speech rate can vary from one message to another, a slot specific speech rate coefficient, provided by the carrier, can also be taken into account.

Two major strategies with respect to duration modelling can be discriminated:

- As the most natural prosody is the one derived from human speech, the possibility is offered to feed the duration module with phonetic transcriptions enriched with duration information copied from natural speech. When customising the MTS system, an argument dictionary containing this information can be built off-line by making use of the prosody transplantation tools (see section 2). If transplanted durations are available in the argument, they are taken over by the duration module and only modified in specific cases — e.g. change a duration in order to cope with a phenomenon such as final lengthening.
- For arguments without transplanted durations, a general purpose duration module is activated. It consists of a cascade of different duration models each having a decreasing specificity. Specific duration models exist for particular arguments such as numbers or date and time indications. The general purpose model is only used if no more specific model is available. Special tools have been developed to speed up the creation of general and special purpose duration models.

3.2.2 Intonation module

The results after duration modelling are input to the intonation module, which produces phonetic transcriptions describing both duration and intonation. After assimilation has been taken care of, the resulting EPT for the argument can be inserted without any further action into the EPT of the carrier.

For each argument, the intonation module calculates a piecewise linear intonation contour based on slot specific intonation models. The slot specific information, provided by the carrier, that can be taken into account is among others the begin pitch, the end pitch, the declination rate and the intonation context (final fall, continuation rise, etc.) of the argument.

4 Discussion

The MTS described above is realised in the context of a dialogue system that places a heavy burden on its hardware environment. It produces high quality speech while morpho-syntactic variations are taken

into account. More specifically, as the MUs and underlying carriers take arguments, it is possible to generate several variants of the same basic message.

- variations on the level of a carrier slot can be paradigmatic: a message ranges over all the elements belonging to a certain semantic category (e.g. product name, cardinality) but the actual message is not known on beforehand.
- variations on the level of a carrier slot can be merely syntagmatic: agreement of all kinds, liaison, contraction, etcetera .
- variations on the carrier level combine both: a message unit can be expressed by other carriers possibly implying other paradigmatic combinations and/or different syntagmatic variations (e.g X replaces Y $\rightarrow$ Y is substituted with X).
- variations on the level of the message units can be semantic: new combinations of message units lead to the creation of new messages.

Also, the MUs and the underlying carriers can be re-used to compose new messages without any loss in speech quality. Good prosody for the carriers is obtained thanks to the prosody transplantation technique. For the slot arguments the same technique can be applied , or prosody is calculated on basis of specific duration and intonation models.

In addition, as the language and task independent core engine is very strictly separated from the language dependent knowledge bases, it is very easy to tailor the MTS system to specific tasks.

References

Ronald Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors. 1995. *Survey of the State of the Art in Human Language Technology.* Cambridge University Press (in press).

Hercules Dalianis. 1996. *Concise Natural Language Generation from Formal Specifications.* Ph.D. thesis, The Royal Institute of Technology and Stockholm University, Department of Computer and Systems Science, Stockholm, Sweden.

Michael Elhadad. 1992. *Using argumentation to control lexical choice: A functional unification-based approach.* Ph.D. thesis, Computer Science Department, Columbia University.

Eduard Hovy. 1995. Overview. In Cole et al. (Cole et al., 1995), pages 161 - 169.

Ehud Reiter, Chris Mellish, and John Levine. 1995. Automatic generation of techical documentation. *Applied Artificial Intelligence*, 9(3):259–287.

Ehud Reiter. 1995. NLG vs. templates. In *Proceedings of the European NLG Workshop 95*, pages 95 - 106.

Harold Somers, Bill Black, Joakim Nivre, Torbj on Lager, Annarosa Multari, Luca Gilardoni, , Jeremy Ellman, and Alex Rogers. 1997. Multilingual generation and summarization of job adverts: the TREE project. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 269 - 276, Washington D.C. Morgan Kaufmann Publishers.

B. Van Coile, L. Van Tichelen, A. Vorstermans, J.W. Jang, and M. Staessen. 1994. Protran: A prosody transplantation tool for Text-to-Speech applications. In *Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP-94)*, pages 423–426, Yokohama, Japan.

B. Van Coile, H. Rühl, L. Vogten, M. Thoone, S. Goss, D. Delaey, E. Moons, J. Terken, J. de Pijper, M. Kugler, P. Kaufholz, R. Krüger, S. Leys, and S. Willems. 1995. Speech synthesis for the new pan-european traffic message control system RDS-TMC. In *Proceedings of Eurospeech 1995*, pages 145–148.

K. van Deemter and J. Odijk. 1997. Context modeling and the generation of spoken discourse. *Speech Communication*, 21:101 - 121.

K. van Deemter, J. Landsbergen, R. Leermakers, and J. Odijk. 1994. Generation of spoken monologues by means of templates. In L. Boves and A. Nijholt, editors, *Proceedings of the Eight Twente Workshop on Language Technology*, pages 87 - 96, Twente.