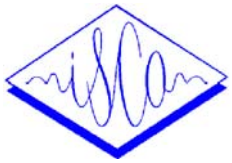

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

INTERACTIVE, TTS SUPPORTED SPEECH MESSAGE COMPOSER FOR LARGE, LIMITED VOCABULARY, BUT OPEN INFORMATION SYSTEMS

Gábor Olaszy, Géza Németh, Péter Olaszi, Géza Gordos*
*Phonetics Laboratory, Institute of Linguistics of the Hungarian Academy of Sciences
Department of Telecommunications and Telematics, Technical University of Budapest
{olaszy, nemeth, olaszi, gordos}@ttt-202.ttt.bme.hu

ABSTRACT

In limited vocabulary speaking systems where several pre-recorded speech items are concatenated to generate the message one of the main problems is to **add new items** to the message set recorded formerly. This procedure is complicated, expensive and the new message(s) cannot be integrated into the original system without leaving the impression that the new items were generated during a later process. In most cases the new items have different tempo, voice timbre and fundamental frequency. Application users would like to have tools which eliminate these problems. To meet these requirements a method and a tool (for Hungarian) has been developed which gives the possibility to create limited vocabulary but open systems. Using this method new speech items can be generated (without recording) with the same voice as that of the original system, so the user will not feel that new message items are combined with the originally recorded items.

Keywords: speech message composer, TTS in background, prosody corrector

1. INTRODUCTION

Speaking information systems with large, limited vocabulary use the method of concatenation of pre-recorded speech items to create the message (station announcers, industrial process control systems, database readers etc.). Two main problems may arise in these systems when **new speech items** are added to the formerly recorded message set.

The first problem is, that adding new message items by traditional practice is very complicated, expensive, long lasting and tiresome (e.g. organisation of the recording with the announcer, making new recording, hire the studio and the equipment, processing the recorded items, embedding and adaptation of the new messages to the original system).

The second problem is that the voice of the newly recorded item -- in most cases -- will not have the same character (in amplitude, tempo, intonation, timbre) as the message items of the original system. The human voice production (timber, speed, accentuation level etc.) changes even for the same person, so the new recording will have other voice characteristics than the original. Therefore, when concatenating these new items with the original ones, incorrect sounding (mainly in fundamental frequency, speed and amplitude) will be heard (Pijper 1997) which is very disturbing to listen to.

One possible solution for this problem is to use a special TTS technology. Such a solution has been reported by Coile et al. (1995) where a prosody transplantation method was combined with TTS to generate high quality speech messages in limited vocabulary systems. The idea of this solution was to copy the intonation and duration values from recorded donor speaker's voice to a TTS. The advantage of this method was that the voice character of the messages of the system was determined by the TTS, so any new message could be generated on the same voice character. The disadvantage of this method was the need of recording for getting the information about the sound durations and the intonation pattern of the new message.

In the present solution the main goal is to eliminate the need of recording. Application users, operators would like to have computer based tools which enable the creation of new message items with the same voice quality as that of the original set without making new recordings.

To meet these requirements a method and a computer based tool (spoken message creator with special TTS in background) has been developed for Hungarian, which gives the possibility to create limited vocabulary but open systems without additional recording. In these systems there is no limitation for the number of message items, because the user, or operator of the system can create any new message item with the same voice quality as that of the original items.

2. THE METHOD

The main characteristics of this new method are: (i) the predefined speech message units (basic vocabulary BV) are recorded directly by the announcer, the new messages (to be created at a later time) are basically created by the special TTS which is designed to work also with the voice of the announcer, (ii) the items of the basic vocabulary will be supplied by predetermined intonation patterns of the intonation rule system. This means that to a certain extent these originally recorded message items will have synthetic intonation instead of their original pitch patterns. (iii) the creation of new message items will be done by the special TTS and the synthesised item will be manually tuned (in duration and amplitude) by a trained person using the prosody corrector. (iv) Finally the new item will be supplied by predetermined intonation patterns of the

intonation rule system. Thus the voice of the new message will have the same voice characteristics and also the same speed, amplitude, and intonation features as the items of the basic vocabulary. This means that in case of concatenation practically no difference will be heard in the structure of new and original message items. With this solution high quality speech message items can be generated in a semi-automatic way and added to the basic message item inventory.

3. PREPARATION OF THE BASIC VOCABULARY

The first step is the determination of the message items of the basic vocabulary. In our case a railway station announcer system was the experimental subject where only declarative forms occurred. The elements of the basic vocabulary (about one hundred items) contained information parts, names, time components and other elements (*train leaves, train arrives from..., at..., fast train arrives, Budapest, West, etc.*). As the first step, the recording of these message items as the elements of the basic vocabulary was made with the selected speaker. A text corpus was designed where these message items were embedded into a proper context for recording. The recorded wave files were labelled by pitch period and sound boundary markers The intonation structure of these message elements was analysed. The main invariant pitch contours for the whole system were defined and summerised in the so-called intonation function set. This function set is used during later processing. This means on one hand, that the message items of the basic vocabulary will be generated according to these contours i.e. they will get a governed synthetic intonation (which is very similar to their original one), on the other hand, these intonation patterns will be used in the special TTS system as well. This common intonation structure works as a bridge between the originally recorded and later constructed items, i.e. the new items will fit to the intonation structure of the items of the basic vocabulary.

4. THE INTONATION FUNCTION SET

The parameters used in this intonation system are: the start pitch value (SP) in Hz, phrase intonation patterns (PhI) like tone groups -- where the change is given in % of SP-- and finally a three level word intonation (WI) structure (for marking words with accented, neutral and negative accented markers). These data are derived from the results of the analysis of real (spoken) messages and the items of the basic vocabulary.

The following phrase intonation patterns were defined

(1). Falling (100% --> 95%) for determination

(2). Falling (100% --> 80%) for final position

(3). Neutral (95% --> 95%) for internal position

(4) Rising (95% --> 100%) for special internal position (e.g. enumeration)

The PhI patterns are used as baselines, the word level intonation patterns are superimposed on the baselines and the calculation of their Hz values is performed from the actual Hz value of the baseline. Thus a range reduction is realised.

The word level accent patterns are as follow:

Accented word (marked by +): the accented syllable gets the following pitch pattern:

beginning point of the vowel	end point of the vowel	end point of the next vowel
100%	115%	100%

The % values are calculated from the Hz values of the actual point of the baseline.

Neutral word: no change in F0 (only the baseline is present)

Negative accented word : 95% --> 95% of the Hz value of the actual baseline.

Pitch generation takes place in two steps: first the baseline is calculated, then the word level pitch changes are realised using the Hz data of the baseline.

By calculation of the Hz values of the baselines the data corresponding to the given % values are calculated from the start pitch value. The start pitch value is the basic fixed point for all pitch calculation. SP can be determined in a configuration file. The value of SP determines the final Hz values in the baseline patterns. Therefore it is important to give the proper start pitch value before the items of the basic vocabulary will be processed with synthetic intonation. Our philosophy was not to change very much the original SP when superimposing the synthetic intonation. Therefore the start pitch in this system was adjusted about 8% higher than the average pith value of the building elements of the TTS. For example if the speakers voice had 105 Hz in the database of the TTS, the SP value was adjusted to 112 Hz in the configuration file. So the higher pitch parts and the lower pitch parts were balanced in the synthetic intonation patterns i.e. the main part of the message had a pitch close to 105 Hz.

5. THE SPECIAL TTS WITH THE VOICE OF THE ORIGINAL ANNOUNCER

The speciality of this TTS system can be summarised in three points. The first is that the system is based on the voice of the very same selected speaker whose voice was used when recording the items of the basic vocabulary. The second is that the inventory of acoustic units contains CVC, CC, VV, CV and VC elements. The third is that the system is embedded into a development tool which contains the prosody corrector as a client. The server is the TTS (Németh et al. 1997).

The inventory of the acoustic units for TTS

In order to avoid spectral discontinuities the main part of the inventory is represented by the set of CVC elements. As vowels represent the main acoustic part of Hungarian we decided to keep the original spectral shape of vowels between consonants (not to cut the vowel in its middle point as it is done by diphone based systems). Therefore the cut points in CVC elements are defined at the middle of the surrounding consonants. The diphone set is generated in a traditional way. During text-to sound

conversion the algorithm analyses the text and uses CVC elements at all possible points. Diphones are used only in those cases where CVC cannot be applied (e.g. CC clusters, VV combinations). The elements of this complex inventory are labelled and supplied by pitch synchronous markers (PIM).
The theoretical number of elements needed for a database for Hungarian (for 14 vowels and 24 consonants) is 8864 CVC elements and 1444 diphone elements.
The creation of this combined inventory represents a very important stage of development. Three syllable meaningless text items (like *aboba, abosa, aboka, akola* etc.) were developed for the recording. Special programs help to cut the necessary parts from the recorded items, to mark sound and pitch boundaries in the elements. Corrections (if necessary) can be done in amplitude and in timing manually by the prosody corrector. CV and VC diphone elements are recorded from another similarly prepared text material.
Using this complex element inventory for TTS, the speech will have a very correct spectral structure which results in high (segmental level) speech quality. This speech will be the body for message creation.

The input form

The basic version of the new message is generated from marked text typed in by the operator. The PhI intonation markers (//x) and the word stress markers (+) can be placed in the text (e.g. //1+*Sebesvonat indul*). This solution gives the opportunity for the operator to determine the basic intonation of the new message. More detailed corrections can be performed by using the functions of the prosody corrector. Setting the final prosody of the new message can be characterised by the correction--listening--correction method.

7. THE PROSODY CORRECTOR

The prosody corrector is an interactive audio-visual tool by which sound duration, amplitude and pitch changes can be precisely performed
This tool helps in making the final form of a new message i.e. corrections in sound duration; amplitude and intonation can be done. The operator can tune the new message in amplitude and in pitch to the structure of formerly created message items. The new message item can be integrated into a full message, which contains earlier created elements. The quality of the generated prosody structure can be evaluated using the listening function.

The prosody matrix

To make prosody manipulations possible a parallel vector with the waveform was created to store and show prosody information in a numerical form This vector appears on the screen as a prosody matrix where the rows represent the prosody data and the columns represent the sounds of the message as it can be seen on Fig 1. This prosody matrix is an advanced version of an earlier one, which was developed for a multimodal German information system (Olaszy et al. 1992).
The sounds in the matrix are represented by corresponding characters and also by sound code numbers. The position row shows the place of the pitch break point inside the sound in %. In the example it is shown how an accent is represented in the matrix (W). The pitch in accented vowel will increase to 115% till the end of the sound and it will decrease till the end of the next vowel. The data in the prosody matrix can be changed manually and these changes are transformed into the wave form representation of the given item immediately. So any kind of prosody content can be generated, i.e. the built in rules can be changed at will. Listening to the result is always possible. So prosody can be tuned effectively.
An example for the intonation structure of a concatenated message as shown by the prosody corrector can be seen in Fig. 1. The sentence is: *Sebesvonat indul Budapest Keleti pályaudvarra három óra ötven perckor, kettõ perc múlva a második vágányról*. (A fast train leaves to Budapest West station at 3.50, within two minutes from platform 2.) The message is concatenated from 7 message items that are marked on the waveform picture by vertical lines. The upper part of the figure contains the beginning part of the real prosody matrix. The window in the middle section shows the wave form of the whole sentence and the bottom part contains the final pitch data.

8.RESULTS AND CONCLUSIONS

The experimental speech message composer tool makes it possible to create new messages in good quality in a railway announcement system. Informal listening tests were carried out with 8 sentences in which new message items and original items were concatenated when forming the sentences. The test subjects could not mark definitely the new items. The advantage of this solution is that there is no need to record the new message items, only a TTS and an audio-visual prosody corrector system is used to create a new message. The disadvantage of this solution is that the operator who creates the new message items has to be skilled in prosody at a certain level. Another disadvantage is that the design and realisation of such tools needs much time and work. The system can be used in all limited vocabulary solutions where the speech material has to be changed sometimes. Such solutions are public transport information systems, industrial process control systems and language teaching interactive software where different speech material has to be formed from time to time.

Acknowledgements

This research was supported by the Hungarian National Committee for Technological Development (OMFB No.96-97-47-1340)

Wave file representation

Three wave files are enclosed to this paper on the CD-ROM with filenames O009x.wav. The files contain railway station announcements in Hungarian, produced by this system..

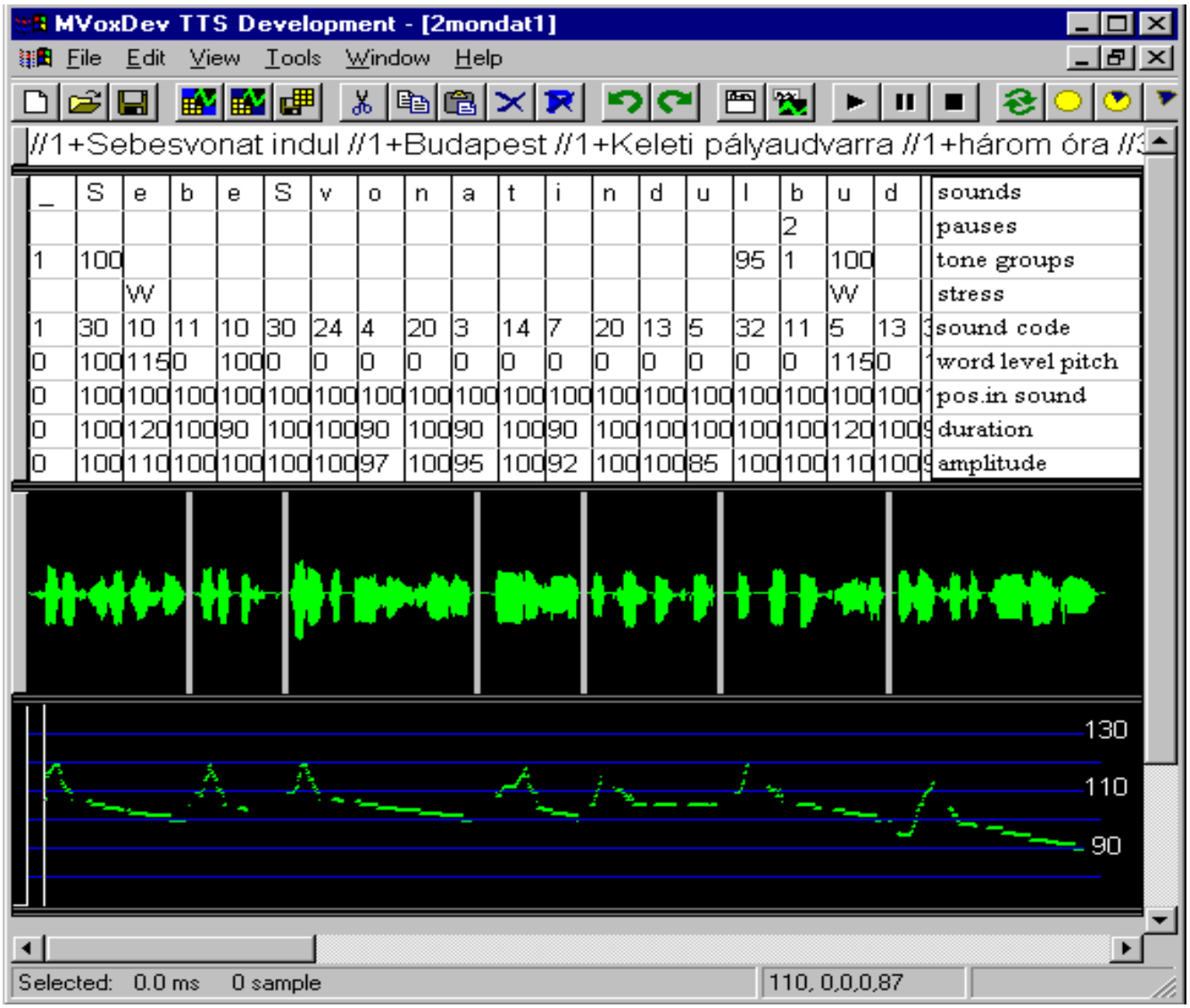

Figure 1
The screen of the message creator showing the data of the sample sentence
(The prosody matrix is shown only partly because of the lack of space in the figure)

References

J. R. de Pijper, High-Quality Message-to-Speech Generation. in a Practical Application. In: J.P. van Santen et al. (eds.) *Progress in Speech Synthesis*. Springer Verlag 1997. pp. 574-589

B. Van Coile, L. Van Tichelen, A. Vorstermans, J. V. Jang, M. Staessen: Protran: A Prosody Transplantation tool for Text-to-Speech Applications, In: Ed. F. J. Lundin (ed.) *Final Report of Cost 233*, Prosodics of Synthetic Speech, Telia, Sweden, 1995. pp. 41-44.

G. Németh, T. Ferenczy, G. Olaszy. Z. Gáspár, A flexible Client-Server Model for Multilingual CTS/TTS Development. *Proceedings. of Eurospeech'97*, volume 5, Sept. 1997, Rhodes, Greece, pp.693-697

G. Olaszy, G. Németh: Prosody generation for German CTS/TTS systems (from theoretical intonation patterns to practical realisation) SPEECH Communication 21. Elsevier Publishers, Amsterdam, 1997, pp. 31- 60.