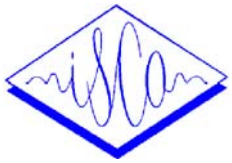

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AN INTERACTIVE TUTORIAL ON TEXT-TO-SPEECH SYNTHESIS FROM DIPHONES IN TIME DOMAIN

Rüdiger Hoffmann, Bettina Ketzmerick[*]*, Ulrich Kordon, Steffen Kürbis*
TU Dresden, Institut für Akustik und Sprachkommunikation, D-01062 Dresden
kom@eakss1.et.tu-dresden.de
[*]BTU Cottbus, Lehrstuhl Kommunikationstechnik, Universitätsplatz 3-4, D-03044 Cottbus
beke@naxos.kt.tu-cottbus.de

10.21437/Eurospeech.1999-164

ABSTRACT

We are presenting an interactive course on speech synthesis which is designed to support the education in speech communication. In the basic section, the fundamental principles of speech synthesis are explained. To explore a complete text-to-speech (TTS) system, the user is provided with access to the Dresden Speech Synthesizer DreSS. The user may type any text, and he may observe how the system processes the text from the first linguistic preprocessing until the acoustic synthesis. A further section is devoted to the crucial problem of correct segmentation of the speech elements used for the concatenative synthesis. The user may select his own diphone segments from a given speech data base. The quality of the segments may be evaluated acoustically, and hints are given to avoid errors in cutting. Thus, the user will learn how to select the segments with good quality. The course is written in HTML and Java and is designed for Internet application.

1. INTRODUCTION

The SOCRATES Thematic Network in Speech Communication Sciences developed proposals for a curriculum in Spoken Language Engineering published in 1998 [1]. We adopted especially these recommendations to develop a course in Speech Synthesis. The interactive tutorial which is presented in this paper will support the teaching process in the chapter on TTS systems which make use of diphones (or similar units) in time domain. Apart from an introductory part, the tutorial concentrates to two main aspects:

1. We show the structure of a state-of-the-art TTS system in detail. This means, the user may type in ASCII text. Then, we show the different levels of processing by explaining the real data structures appearing at the interfaces from block to block when the given text is processed. Finally, of course, the user will listen to the synthesized result.

2. We show in detail the most crucial part which concerns preparing and selecting the inventory. To get some experience in this field, the user may choose utterance, may select speech units from the waveform, may listen to them and cut. In this way, he is able to produce a small inventory from which he may combine a selected number of synthetic words. He is able to study the influence of errors which arise when cutting errors are made or unsuited elements are concatenated.

The synthesis system used for demonstration is the Dresden Speech Synthesizer DreSS [2]. For this purpose, the system was broken into its essential components. The data structures at the interfaces of these components have been made visible and accessible in an interactive way. The inventory used is one of the male German voices of DreSS. A future version will make available other voices as well as the multilingual facilities of DreSS by successively adding our English, Russian, Czech, Chinese, and Italian inventories.

A first version of the tutorial was presented at the MATISSE workshop [3]. The version presented now features essentially more material in the explanatory parts as well as a new section for selecting speech segments from some own acoustic input.

To use the tutorial, Netscape 4.xx will be necessary at least. The acoustic output can be in the .wav, .au, or .aiff formats.

2. THE GENERAL PART

The tutorial starts with an introductory part on the background and the usage of the tutorial. Starting from this point, the user has access to a first main part which explains the structure of a TTS system in general. For this purpose, the block diagram of a TTS system is shown (Figure 1).

To obtain more information, the user may click on the different symbols of this flowgraph. Doing so, a window appears which explains the function of the corresponding block or the components of the corresponding database, resp.

The tutorial is restricted to the presentation and concatenation of the speech units in time domain. This means, other (parametric) methods for presenting the speech units will not be covered.

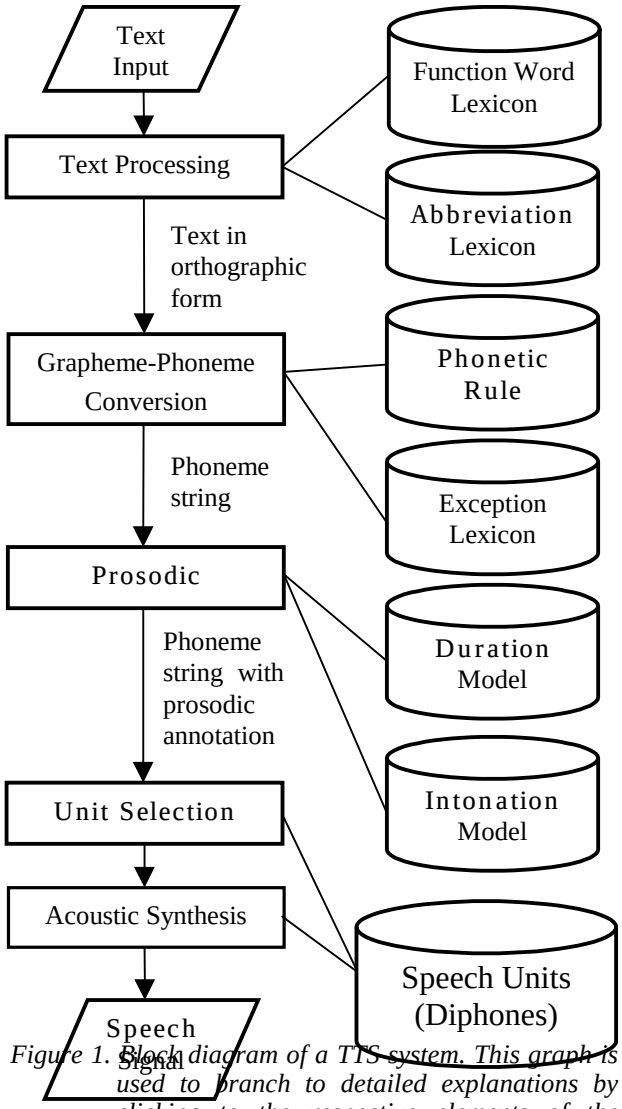

Figure 1. Block diagram of a TTS system. This graph is used to branch to detailed explanations by clicking to the respective elements of the diagram.

3. THE DreSS DEMONSTRATION

In the main part of the system, the user may investigate the behaviour of a complete TTS system by typing any sentence into the input field of our system DreSS (Dresden Speech Synthesizer). E. g., the input could be:

```
Zur Demonstration wird ein Text ausge-
geben.
```

In the simplest case, this text is transmitted to our server, is processed, and the speech output file is sent to the user. However, if the user is interested in the different levels of processing, he is allowed to switch on a step-wise operation which makes accessible the results of the different modules. For this purpose, the viewgraph according to Figure 1 is extended by additional bars which may be clicked onto.

To follow the example given above, the result of the text processing step will be:

```
{Utterance:begin}{ClauseType:final}
#zur#{WClass:Nom}demonstration#
{PhraseBound:b3}wird#ein#{WClass:Nom}t
ext#ausgegeben#*.{Sentence:end}
{Utterance:end}#@@
```

Secondly, the result of grapheme-phoneme conversion will be shown:

```
{Utterance:begin}{ClauseType:final}
{WordAcc:1}tsu:6{WordAcc:0 0 0 2}
dEmOnstratsjo:n{PhraseBound:b3}
{WordAcc:1}vIrt{WordAcc:1}?aIn
{WordAcc:2}tEkst{WordAcc:1 0 0 0}
?aUsg@ge:b@n{Sentence:end}
{Utterance:end}
```

The next levels which may be investigated by the user are the results of duration modelling and of intonation modelling. In both cases, the code shown above will be annotated by the corresponding information. Since the size of these data sets in increasing from level to level, we will nor print it here. The last data which may be observed is the complete information on diphone level which is produced by the Unit Selection module. In our example, it reads as follows:

```
PAUSE | |   0 1   ;
_,43

{Utterance:begin}
{ClauseType:final}
{WordAcc:1}
ts    | *   0 2   ;
t,0 s,72
t,0.99,0.20 s,0.99,0.20

su:   * *   0 2   ;
s,72 u:,99
s,0.99,0.20 u:,0.99,0.20

u:6   * *   0 2   ;
u:,99 6,72
u:,0.99,0.20 6,0.97,0.20

{WordAcc:0 0 0 2}
6d    * |   0 1   ;
6,72 d,0
6,0.97,0.20
<etc.>
```

Of course, the SAMPA notation which is used here as well as the additional information in these data will be explained.

Furthermore, the user may evaluate the effect of different strategies for modelling the prosodic components. For the duration as well as for the intonation, two models are implemented which may be selected by clicking a button.

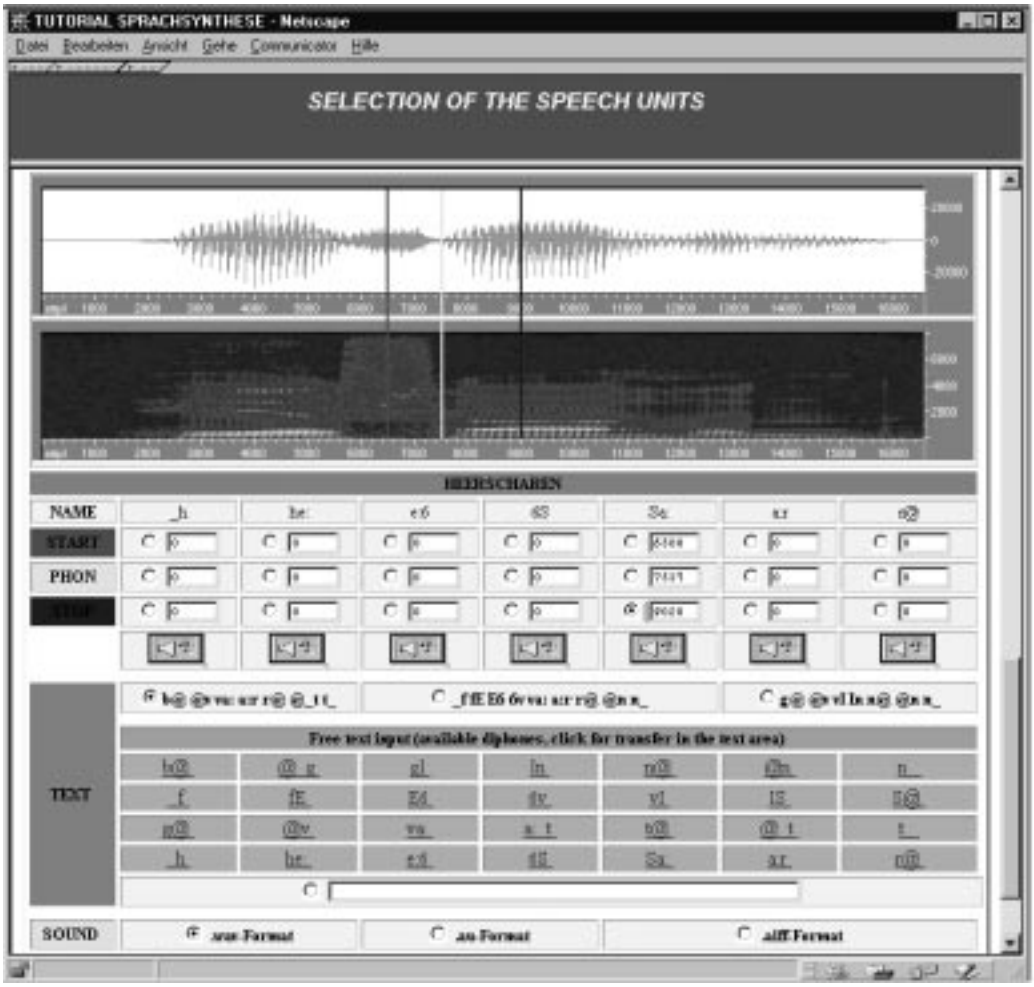

Figure 2. Example for the selection of a diphone. In the upper part, the waveform and the spectrogram of one of the predefined words (the German word ‘Heerscharen’*) are shown. By means of the buttons in the medium part, the user has selected the boundaries of the diphone* Sa: *(SAMPA notation). By means of the lower part, the user may finally combine some of the selected diphones to produce a new word or logatome, resp., for evaluating the quality of the selection process.*

Duration models available:

- a linear model which is similar to Klatt [4],
- a data driven model which calculates the durations of the phrase, of the syllables, and of the sounds in three layers [5].

Intonation models available:

- a linear model which combines a simple declination line with piecewise linear accent information,
- an adaptation of the well-known Fujisaki model [6] to German which was elaborated by Mixdorff [7].

Experience shows that the simpler models are preferred for the synthesis of short statements while the more sophisticated models are suited for the synthesis of longer phrases.

4. THE SPECIAL CHAPTER ON SEGMENTATION

4.1 Purpose

The quality of the speech synthesized by a TTS system depends stongly on the experience of the people who produce the speech segments [9, 10]. That’s why, we added a special chapter to the tutorial in which the user will be trained in selecting speech segments (especially diphones) from speech material (e. g., words). This chapter is subdivided into a descriptive section (text and illustrations) and two experimental sections. In the text, the following topics are covered (for details, cf. [3]):

- Sound elements for speech synthesis
- Requirements for the speech material
- Manual segmentation
- Support by automatic segmentation
- Characteristics of the sound classes
- Conventions
- Problems

4.2. Experiment I

The user may get his own experiences in segmentation by producing a selected set of own diphones. In Table 1, we present four words which are available to the user to select his diphones. These utterances are spoken by the standard speaker of the DreSS system.

The procedure for selecting diphones from the words given in Table 1 is supported by a graphic representation of each of the four words in time and frequency domain (Figure 2). If the user wants to select a diphone, he may use a cursor to indicate

Table 1. Word inventory for individual diphone selection

Word	Diphone sequence (SAMPA)
gewatet	g@ @v va: a:_t t@ @_t t_
verwischen	_f fE E6 6v vI IS S@ @n n_
beginnen	b@ @_g gI In n@ @n n_
Heerscharen	_h he: e:6 6S Sa: a:r r@ @n n_

- the start point of the diphone (START),
- the end point of the diphone (STOP),
- the boundary between the two sounds which form the diphone (PHON).

If the diphones are prepared in this way, the user may press a "send" button, and the complete information will be transferred to the host where the database is changed according to the selected values. Diphones which were not influenced by the user make use of predefined boundaries.

From these diphones, the user may synthesize new words or logatomes to evaluate the quality of his selection. The synthesis proceeds by remote access to DreSS. To avoid additional influences of the grapheme-phoneme conversion and other parts of the TTS system, the strings to be synthesized have to be written in SAMPA.

4.3. Experiment II

As a further interactive part of the tutorial, a section for cutting the own speech material is offered to the user. This experiment runs locally at the computer of the user. He may speak some utterances into his microphone. The speech signal is visualized, speech segments may be cut, and combinations of the speech segments are synthesized if wanted for evaluating the quality of the segments again.

5. CONCLUSION

We presented the recent version of our tutorial on speech synthesis. Further work will cover more details in preprocessing and prosodic manipulation. Furthermore, additional voices and languages will be included.

Access to the tutorial is via the internet address

www.ias.et.tu-dresden.de/kom/lehre.

6. REFERENCES

[1] Bloothoft, G., et al. (1998). *The landscape of future education. Part 2: Proposals*. Utrecht: Led 1998.

[2] Hoffmann, R., D. Hirschfeld, O. Jokisch, U. Kordon, H. Mixdorff and D. Mehnert (1999), Evaluation of a multilingual TTS system with respect to the prosodic quality. *Proc. XIVth Int. Congress of Phonetic Sciences, San Francisco*, August 1-7, 1999.

[3] Hoffmann, R., U. Kordon, S. Kürbis, B. Ketzmerick and K. Fellbaum (1999), An interactive course on speech synthesis. *Proc. ESCA/SOCRATES Workshop MATISSE, London*, April 16-17, 1999, pp. 61 – 64.

[4] Klatt, D. H. (1979), Synthesis by rule of segmental durations in English sentences. *Frontiers of Speech Communication Research*, ed. by B. Lindblom and S. Öhman, Academic Press, London, pp. 287 – 299.

[5] Jokisch, O., D. Hirschfeld, M. Eichner and R. Hoffmann (1998), Multi-level rhythm control for speech synthesis using hybrid data driven and rule-based approaches. *Proc. ICSLP '98*, Nov/Dec 1998, Sydney, pp. 607 – 610.

[6] Fujisaki, H. (1997), Modeling the process of fundamental frequency control of speech for synthesis of tonal features of various languages. *1997 China-Japan Symposium on Advanced Information Technology*, Invited Plenary Lecture, March 1997, Anhui, pp. 1 - 12.

[8] Mixdorff, H. (1998), *Intonational patterns of German – Quantitative analysis and synthesis of F0 contours.* PhD thesis, Dresden University of Technology.

[9] Ketzmerick, B., and K. Fellbaum (1997), Zur Hörbarkeit von Segmentierungs- und Konkatenationsfehlern bei der Erstellung einer Lautelemente-Bibliothek für die Sprachsynthese. *Proc. DAGA 97*, March 1997, Kiel, pp. 539 – 540.

[10] Ketzmerick, B. (1997), Segmentierungs- und Konkatenationsprobleme bei der Erstellung einer Lautelemente-Bibliothek für ein Sprachsynthesesystem. *Proc. ESSV 97*, Aug 1997, Cottbus, pp. 256 – 263.