FELIX - A TTS SYSTEM WITH IMPROVED PRE-PROCESSING AND SOURCE SIGNAL GENERATION

Georg Fries and Antje Wirth
Deutsche Telekom Berkom GmbH
Forschungsgruppe Sprachverarbeitung
Am Kavalleriesand 3, D-64295 Darmstadt, Germany
E-mail: {friesg, wirth}@tzd.telekom.de

ABSTRACT

Felix is our recent PC-based TTS research-system for testing, analyzing, and evaluating TTS algorithms. The object-oriented interface allows efficient algorithm improvement and overall system prototyping by combining different modules. The results of each TTS-processing step can be monitored and all kinds of data may be reviewed and modified. The paper will outline the algorithms currently implemented in the Felix system, focusing on lexical analysis, duration modeling, and source signal generation, where we suggest ways to improve intelligibility and naturalness of synthetic speech.

1 INTRODUCTION

In general the system architecture is the same in most TTS-systems: a sequential process of text pre-processing, prosody generation and signal production. Within our current Felix environment we built up such a classical TTS system with the following modules: text normalization, lexical and syntactical analysis, duration and intonation modeling, and hybrid signal synthesis with phoneme-specific source generation. In Felix, the only link between these modules is a data structure that serves as an information container and that is the backbone of the system (Fig. 1). This data structure, designed according to the structure of speech, is used to store all property values, any module obtains from its processing, for later use by other modules. Hence it is possible to combine modules in a flexible way and concentrate on the core of an algorithm. Because of its object-oriented design, the container can easily be expanded.

The overall quality of synthetic speech is fairly good. Nevertheless, some studies [1] show, that humans tend to reject synthetic voices. The intelligibility of synthetic speech is in part reduced by problems arising from pre-processing of the text, like pronunciation errors, errors caused by miss-interpretation of acronyms, or by spelling mistakes. The lack of naturalness can be addressed by increasing the quality of prosody and signal generation. Both intelligibility and naturalness, may be improved by adding semantic processing to TTS-systems. Further, many applications require speech generation capabilities instead of reading a given text. That means to generate utterances from tasks like „say hello“.

These examples show, that the problem of TTS has been shifted from finding an overall solution to optimizing and adjusting parts of it, in order to increase the quality of synthetic speech and to fit to the needs of special applications. Felix supports this by its modular architecture and its sophisticated data structure.

In this paper we give an overview of the modules currently implemented in Felix, focusing on lexical analysis, duration modeling, and source signal generation, where we suggest ways to improve intelligibility and naturalness of synthetic speech.

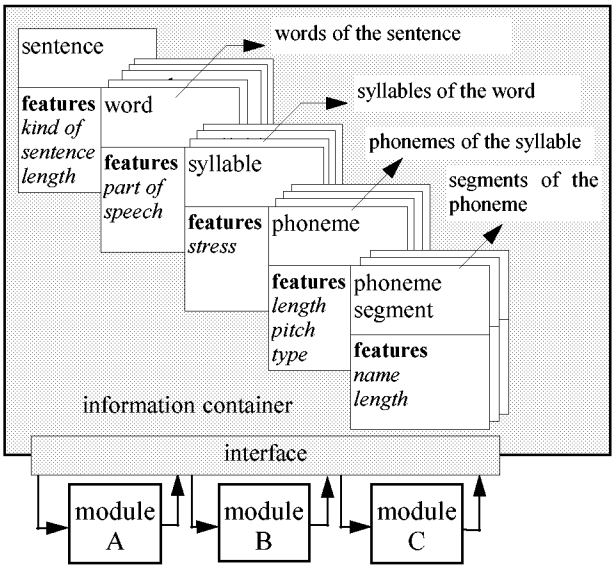

Fig. 1 Scheme of the Felix architecture

2 TEXT-PRE-PROCESSING AND PROSODY GENERATION

In Felix, pre-processing of text and prosody generation are currently done sentence-by-sentence in a classical sequence of steps:

- Text-normalization is based on a simple pattern matching algorithm and a table lookup in a list of acronyms. Information, like the sequence of words, is determined and written to the information container. If transcriptions are available, e.g. from the lists of acronyms, they can directly be stored in the container, skipping lexical analysis. The lexical analysis is described in chapter 2.1.
- A simple syntactical analysis is done by an algorithm of Zingle [2], where a window of three words is shifted over the sentence word by word. Zingle's algorithm is based on the theory of group accent. Phrasing and accentuation of a sentence are determined on the basis of the sequence of parts of speech in the window.
- Estimation of phoneme duration is described in section 2.2.
- The fundamental frequency control algorithm is based on a model proposed by Adriaens [3]. In this model the fundamental frequency contour of a sentence is built by superposition of a set of declination lines and a sequence of complex F_0 contours, which result from a rule-based combination of linear atomic F_0 movements.

2.1 Lexical analysis

The lexical analysis is based on a combination of several algorithms. Its output consists of word transcription, syllabification, accentuation, morphological structure, and part of speech. Its core is constituted by a morpheme lexicon and a fast search algorithm.

The morpheme lexicon is kept in a relational database for easy maintenance of the data. It contains information about transcription, stress, and syllabification of basic constituents (e.g. stems, affixes, linking letters, inflectional endings and their possible combinations). In addition, uninflected words and compounds (consisting of up to four stems) are included in the lexicon as links between the stems and their affiliated constituents. Additional information, for instance, the type and the way of declension or conjugation and the part of speech was added to these complex constituents. Table 1 gives an impression of the current lexicon size.

constituent	number
basic	
stems	6 725
affixes	910
inflectional endings and linking letters	53
complex	
stems with affixes	15 816
compounds	15 364
function words, frequent words	1 240

Table 1 Size of the morpheme lexicon

The fast search algorithm is a combination of hashing, intelligent buffering, and binary search. On a PC 486 under MS Windows 3.11, 100 000 searches in a lexicon of approximately 150 000 entries take on average 2.5 msec.

The process of lexical analysis starts with a search for complete words in several lexical resources, e.g. a list of frequently used words, a lexicon of proper names [4] and a full form lexicon of common words. Which lexical resources are used for the full form search and the order of lookup can be configured to adjust the lexical analysis to special applications. The list of frequently used words is obtained from the morpheme lexicon. The lexicon of full forms is automatically generated from the complex constituents included in the morpheme lexicon. In this way, lexical analysis can be optimized concerning time and transcription errors, since no parsing is necessary for all words included in the full form lexicon. In addition, maintaining a morpheme lexicon is easier, than a full form lexicon, where every inflection has to be processed separately. Words, that are still unknown, are analyzed by a morpheme parser (similar to [5]) in the next step. The parsing strategy is based on the following syntax of German words (EBNF notation):

word={prefix}stem{suffix}[[linking letter]word][inflectional ending]

The parser uses the basic constituents of the morpheme lexicon for morpheme analysis as well as the information about the complex constituents for disambiguation and time optimization. The morpheme analysis of a word results in a list of constituents. The transcriptions of the constituents have to be assembled in a post-processing step, using a small set of rules, in order to get the right stress and syllabification. If a word could not completely be analyzed, the parser supplies prefixes, suffixes and an inflectional ending. The rest of the word is passed to a simple rule transcription algorithm.

2.2 Duration modeling

The duration of phonemes is influenced by several articulatory and linguistic quantities. It is nearly impossible to determine the influence of one single quantity separately. For that reason, duration measurements of several realizations of a phoneme, supposed to be influenced by one special condition (e.g. positioned in an accented syllable) often result in duration values, that seem to be quite stochastically. The main idea of our duration model is to take into account some of the most significant durational influences only, while modeling the others by a random process.

2.2.1 Basic parameters

The durational behavior is characterized by a triplet of basic parameters <*a, m, b*>, consisting of a phoneme duration range <*a, b*> and a mean duration value *m*, representing the intrinsic duration of a phoneme. A set of parameter triplets was obtained for every phoneme from measurements of minimum, maximum and medium duration values in fluently spoken speech [6]. Each triplet within a set represents the durational behavior of one phoneme due to one complex of features. It is considered that phoneme duration varies between <*a,b*> based on a Gaussian random process.

2.2.2 Durational influences

Based on the results shown in [6], the following influences are taken into account for the duration model:

- **accentuation** with the feature values *sentence focus*, *group accent*, *word accent* and *unstressed*
- **group position** of the syllable with the feature values *final* and *not final*
- **syllabic position** of a phoneme with the feature values *onset*, *nucleus* and *coda*
- **phoneme** including all German phonemes, the glottal stop and a pause
- **speed** of speech with the feature values *slow*, *normal*, *fast*

Sets of parameter triplets were estimated for the feature complexes <accentuation, syllabic position, speed> and <group position, syllabic position, speed> for every phoneme.

2.2.3 Calculation process

The input information of the model consists of the phoneme, its feature values and the speed value (e.g. <t, onset, unstressed, final, 0.9>). The parameter triplets for slow, normal and fast speed are chosen from the set of the phoneme, according to the feature complex, estimated from the feature values without consideration of the speed value. The speed is adjusted by calculating a new triplet from the chosen ones in the way that the new values for <a,m,b> are gained from the linear interpolation between the corresponding basic parameter values of the slow and normal or the normal and fast triplets, respectively. The output phoneme duration is randomly selected from this newly calculated triplet.

Due to observations from [6], the overall variation of duration is phoneme-specifically limited to the range between the lowest and the highest measured value. Speech, that is faster or slower than the speech resulting from these values, is produced by further lengthening or shortening of the pauses only. This is based on the observation, that there is a strong correlation between the duration and occurrence of breath pauses and the speed of speech - utterances, spoken slowly, contain on average more and longer pauses than the fast spoken ones.

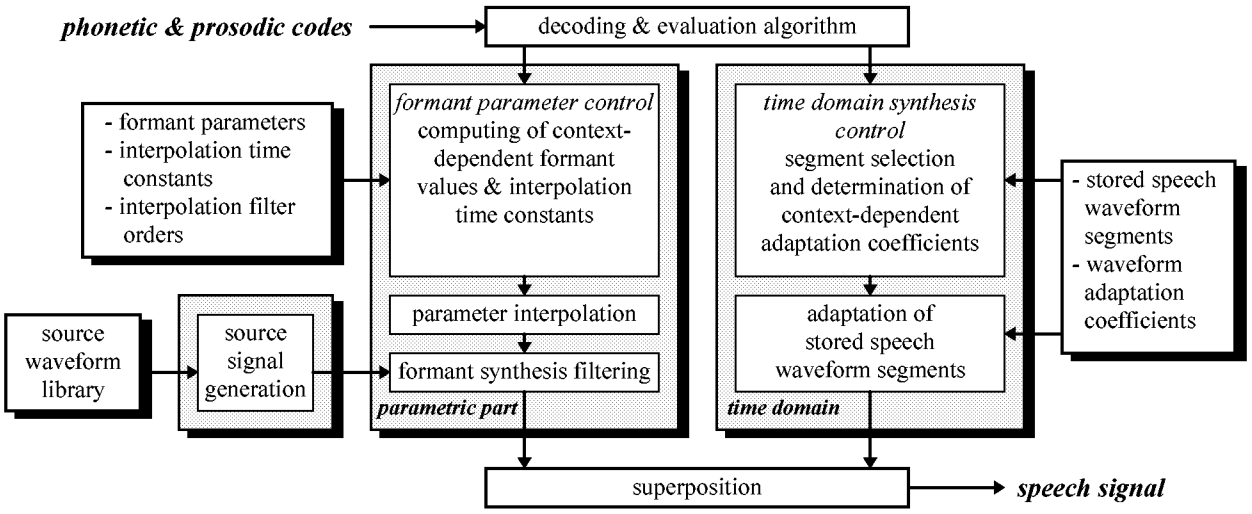

Fig. 2 Hybrid speech signal generation stage of the Felix system

3 SPEECH SIGNAL GENERATION

3.1 Hybrid synthesis approach

Fig. 2 shows the signal generation stage in the Felix system. We apply a hybrid signal synthesis approach that combines speech synthesis in the time and frequency domain, as first introduced in [7]. A decoding and evaluation algorithm processes the input code sequence and determines, in a phoneme-dependent way, the subsequent alternative or simultaneous use of the parametric system part and the time-domain part. Vowels and nasals are exclusively produced by the parametric system part - a formant-based synthesizer. Voiceless consonants are entirely generated in the time domain. For this purpose the time-domain component provides and modifies stored speech waveforms that are superimposed on the formant synthesis signal. Since all voiceless sounds are created in the time domain, the formant synthesizer part requires no voiceless excitation. To produce voiced fricatives, voiced plosives, and voiced/voiceless-transitions, both system parts are simultaneously applied. A better quality of fricatives and plosives has been achieved, whereas the flexibility in fundamental frequency variation is preserved. The generation of the voiced source signal is described in the next section.

3.2 Improvement of source signal generation

Many investigations focus on improving naturalness of synthetic speech by applying a sophisticated glottal source excitation [8,9]. Some approaches [10,11] use natural glottal source signals. In [12] we proposed such a source signal generation scheme, which is based on concatenation and modification of phoneme-specific natural source segments. Fig. 3 describes the steps of the source signal composition. First the segment selection unit selects and composes source waveforms as a function of phoneme class, duration, and the requested F_0-contour. These segments have been extracted in an off-line process by an inverse filtering method and stored in the waveform library. The pitch of the selected source segments is always higher than or equal to, the requested pitch. Hence during the synthesis process, the pitch of the selected segment is lowered by extending each single pitch period length to the required value. Finally the generated source signal is smoothed at the segment boundaries. In the current Felix configuration we improved this scheme as follows:

- The length of the stored segments has been extended.
- The source segment extraction process has been modified by introducing representative formant parameters for inverse filtering.

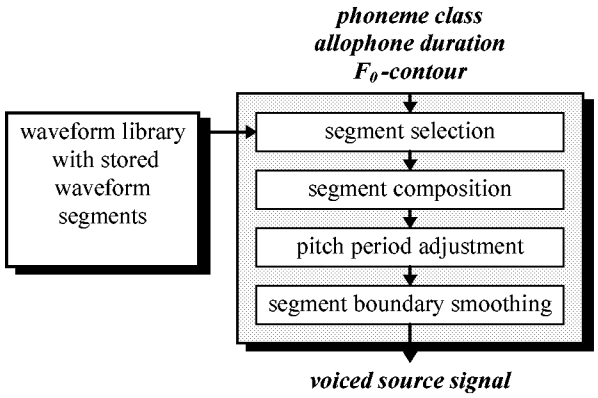

Fig. 3 Generation of the voiced source signal

3.2.1 Extension of the source signal segment size

In our previous source generation scheme the source waveform composition was done by random selection of single, double, or triple source pulses as a function of the desired pitch [12]. Now, we use segments that are at least as long as the current allophone duration. Hence the number of source segment boundaries is significantly reduced. Further, the inherent variations in the excitation signal are preserved within a longer interval.

3.2.2 Source segment extraction using representative formant parameters

To extract the required source signal segments for the waveform library, we recorded 12 allophones at 12-15 pitch levels and analyzed the recordings. Fig. 4 outlines the off-line processing that was applied to each allophone. After automatic formant analysis, a set of formant parameters, that best estimates the spectral properties was interactively determined. These *Representative Formant Parameters* (RFP) were used to model the inverse filter of the following analysis stage. Representative means that these parameters are valid for the whole duration of one allophone during the inverse filtering process. The motivation of using RFP is as follows: In formant synthesis, the spectral shaping of the synthesized speech is done by controlling the formant values by rules. This leads to relatively smooth formant movements, which again cause in part the

unnatural smoothness of a typical formant synthesizer output. Hence, to improve naturalness, it is necessary to model the signal's temporal details more exactly. We can either vary the formant movements frame by frame (as in interactive copy synthesis), or we can add more details to the excitation signal. It is difficult to use rules to model exactly the formant movements, which are obtained by copy synthesis. Hence our approach is to keep the original speech signal's temporal details in the excitation signal, by inverse filtering with *constant* parameters based on the RFP.

In the next stage, the pitch values are measured and adjusted to one exact F_0 value for all periods. Afterwards the boundaries of each single period are marked. The exact pitch and the location of the pitch boundaries are stored as additional parameters in the waveform library.

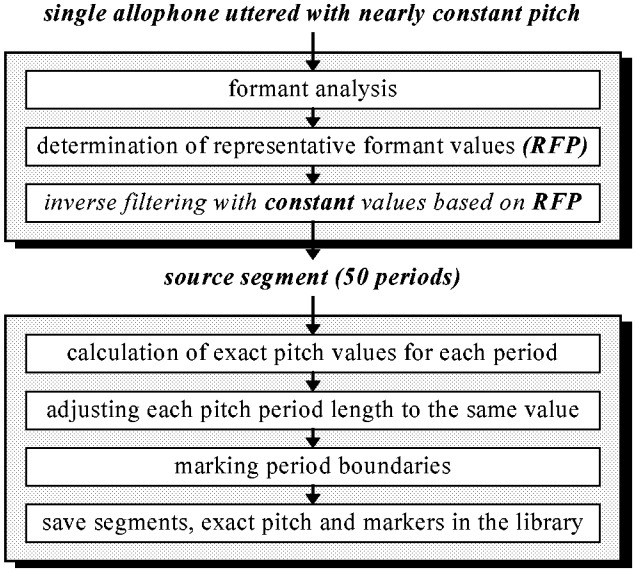

Fig. 4 Off-line extraction of source segments

Informal listening tests with isolated allophones showed that the naturalness of the voiced sounds has been improved compared to the results of our previous scheme, which concatenated shorter source segments. In a future test we will evaluate this result for fluent synthetic speech.

3.2.3 Determination of source segment variants

In Felix the waveform library uses phoneme specific source segments. In order to define the different phoneme classes, we compared the quality of allophones synthesized with source segments, that were extracted from different phonemes.

	/a/	/i/	/o/	/u/	/e/	/m/	/n/	/N/	/l/
res{/a/}	+	-	o	+	o				
res{/i/}	-	+	-	+	-				
res{/o/}	-	o	o	o	o				
res{/u/}	-	-	o	o	-				
res{/e/}	o	o	+	+	+	-	-	-	o
res{/m/}						o	-	o	-
res{/n/}						-	+	o	o
res{/N/}						-	-	+	+
res{/l/}						o	-	o	+

Table 2 Scores (+ good, o medium, - poor) for allophones synthesized with source segments extracted from different phonemes, res{/i/} means: source segment extracted from /i/

Table 2 shows some relative scores, i.e. source segments extracted from the vowel /e/ (res{/e/} in Table 2) show good results if used for synthesizing the vowels /u/, /o/, /e/. In Felix, the waveform library currently has the following structure:

- 6 phoneme-specific sections, each representing a phoneme class: res{/a/}, {/i/}, {/e/}, {/m/}, {/n/}, {/N/}.
- Within each phoneme class, 12 different pitch value sections exist.
- Each pitch value section consists of a source segment with the length of 50 periods.

4 CONCLUSION

Our TTS research-system Felix was introduced. The object-oriented interface allows efficient algorithm improvement and overall system prototyping by combining different modules. We outlined the algorithms currently implemented in the Felix system. We described a way to reduce the rate of pronunciation errors by performing a lexical analysis, based on a morpheme lexicon. The simple duration model of Felix was designed to increase the naturalness of synthetic speech by using random values in the calculation of phoneme duration.

Further in the current Felix configuration we modified the source signal generation by extending the size of the stored source segments in the waveform library and by introducing representative formant parameters (RFP) to the off-line processing. Informal listening tests with isolated allophones show that the naturalness has been improved compared to the results of our previous scheme, that concatenated shorter source segments.

5 REFERENCES

[1] Léwy, N.; Hornstein, T.: *Text-to-Speech Technology: A Survey of German Speech Synthesis Systems.* UBILAB Technical Report 94.10.2, Zürich, 1994

[2] Zinglé, H.: *Traitement de la prosodie allemande dans un systeme de synthese de la parole.* Thèse pour le Doctorat d'Etat, Université de Strasbourg II, 1982

[3] Adriaens, L. M. H.: *Ein Modell deutscher Intonation.* Dissertation, Technische Universität Eindhoven, 1991

[4] ONOMASTICA: *Multi-language pronunciation dictionary of proper names and place names.* Final Report, LRE-61004, 1995

[5] M. Kommenda: *Automatische Wortstrukturanalyse für die akustische Ausgabe von deutschem Text.* Dissertation. Technische Universität Wien, 1991

[6] A. Wirth: *Investigation of the durational structure of German fluently spoken speech.* In COST 233 Prosodics in Synthetic Speech - Final Report, 1995

[7] G. Fries: *Phoneme-dependent Speech Synthesis in the Time and Frequency Domains.* Proc.Eurospeech, pp. 921-924, 1993.

[8] A. E. Rosenberg: *Effect of glottal pulse shape on the quality of natural vowels.* J. Acoust. Soc. Am., vol. 49, pp. 583-590, 1971.

[9] G. Fant, J. Liljencrants, Q.-G. Lin: *A four-parameter model of glottal flow.* Speech Trans. Lab. Q. Prog. Stat. Rep., vol. 4., Royal Institute of Technology, Stockholm, pp. 1-13, 1985.

[10] K. Matsui, S. D. Pearson, K. Hata, T. Kamai: *Improving naturalness in text-to-speech synthesis using natural glottal source.* Proc. ICASSP, pp. 769-772, 1991.

[11] D. M. Howard, A. P. Breen: *Method for dynamic excitation control in parallel formant speech synthesis.* Proc. ICASSP, pp. 215-218, 1989.

[12] G. Fries: *Hybrid Time- and Frequency-Domain Speech Synthesis with extended Glottal Source Generation.* Proc. ICASSP '94.