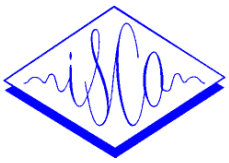

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

HADIFIX - A SPEECH SYNTHESIS SYSTEM FOR GERMAN

Thomas Portele, Birgit Steffan, Rainer Preuß, Walter F. Sendlmeier, Wolfgang Hess

Institut für Kommunikationsforschung und Phonetik der Universität Bonn (IKP)
Poppelsdorfer Allee 47, D-5300 Bonn 1, Germany

ABSTRACT

HADIFIX is a speech synthesis system for German. It transforms a phonemic input string with accent markers into a speech signal. To obtain high-quality speech, time-domain elements with a structure similar to demisyllables are concatenated. A method to generate vowel-to-vowel transitions for time-domain units is used when necessary. Prosodic manipulations are carried out using the TD-PSOLA algorithm [1]. An F_0 contour is generated using Fujisaki's model [2,3]. The quality of the synthesized speech was compared with that of natural speech and of an LPC-based version using a rhyme test, a word test, a dictation, and a subjective ranking test. While the intelligibility is high, the difference in naturalness between synthesized and natural speech is still large. Another outcome of the test is the superiority of the time-domain units, compared with the LP-coded ones.

10.21437/ICSLP.1992-157

1. INTRODUCTION

Since the development of the PSOLA algorithm [1], time-domain units have become popular in speech synthesis. A careful choice of the unit structure is required here, because the possibilities to manipulate time-domain units are limited. For a language like German, where complex consonant clusters are quite frequent, syllable-oriented units appear to be a good choice [4]. To obtain a reasonable size of the unit inventory, syllables are often composed using two demisyllables (DS's) [5,6].

The HADIFIX system [the acronym stands for "*Ha*lbsilbe" (demisyllable), *di*phone, and suf*fix*] was developed to investigate the advantages and difficulties that arise by the use of syllable-oriented time-domain units. This paper describes the unit inventory structure, rules and methods for the concatenation of time-domain units, prosodic manipulations, and the evaluation of the speech generated by some versions of HADIFIX.

2. INVENTORY STRUCTURE

Based on the work by Dettweiler [6], who split the final DS into a rudiment and a suffix before some obstruents, the principle of DS splitting is applied to all final DSs (except those with syllable-final vowels). An inventory of HADIFIX consists of units of three classes [7]:

- initial demisyllables extending from the beginning of a syllable to the beginning of the stationary part of the syllabic nucleus,
- diphonic elements containing the main part of the syllabic nucleus and either the first half of a postvocalic sonorant, or the transition to one of three obstruent classes (labial, alveolar, velar), or the vowel offset for an open syllable,
- suffixes with syllable-final consonant clusters.

About 1080 units (750 initial DSs, 150 diphones, 180 suffixes) are sufficient to synthesize nearly all German words, including those with uncommon sound combinations originating from foreign languages, such as "Szene" (scene) or "Dschungel" (jungle).

Two inventories were recorded: one male and one female voice. Most units (except those containing schwa) were extracted from secondary-stress syllables embedded in carrier sentences [7]. To perform pitch-synchronous manipulations of the signal, a laryngograph was used during the recording for exact marking of pitch periods. The units were cut manually using SONA, a combined sonagraph / waveform editor developed by D. Stock at the IKP [15].

Each inventory requires 5 MB storage (16 kHz sampling frequency, 16-bit words, no data reduction).

In most cases, the reduction in inventory size achieved by the shift from two to three syllable-constituting units has no negative influence on the quality of the resulting speech. Due to allophonic variations, however, problems arise, when a boundary exists inside a postvocalic liquid (especially /r/). A better location of the splitting point in final DSs might be the voiced-unvoiced boundary [7].

3. CONCATENATION

In the HADIFIX system a speech signal is generated by the concatenation of time-domain units. Compared with other coding techniques (LPC or formant coding), the possibilities to manipulate the units are quite limited. In its current state, HADIFIX supports only three operations: cut, smoothing over one period, and spectral interpolation via a transformation to LPC-parameters. Due to the unit structure, where many coarticulative phenomena are implicitly contained in the units, only a few concatenation rules are necessary. For most of these, the two simple operations *cut* and *smooth* are sufficient. Table 1 lists all the concatenation rules applied.

The need for artificial transitions between units (spectral interpolation) is fairly small for a diphone system, because all transitions are included within a unit. But a syllable-oriented unit structure must face problems such as two adjacent vowels at a syllable boundary. Instead of adding additional units, an algorithm was developed because of its greater flexibility [4,7]: The last pitch period of the first unit and the first pitch period of the second unit represent the two target frames. They are LP analyzed. The respective residual signals are calculated, and these are time aligned so that they can be added pitch-synchronously in the next step. Within the interpolation interval, the two LP transfer functions are combined by weighted addition of log area ratios of corresponding order, while the two time aligned residuals are also weighted appropriately and added. Thus an array of LP parameter sets and a corresponding array of residuals are constructed. LP synthesis, where each parameter set is excited by the pertinent residual signal, yields a transition between the two targets. The results are good; the artificial transitions fit between the time-domain units without attracting any unwanted attention.

4. PROSODIC MANIPULATIONS

Prosodic manipulations of the speech signal are carried out using the TD-PSOLA algorithm [1]. This method allows modifying duration, intensity, and fundamental frequency.

The duration structure of an utterance is generated by a set of syllable-level rules [8]. These rules were formulated after the statistical analysis of a small real-speech corpus. A few intensity

- between initial DS and diphone:
  smoothing in the range of one pitch period

- between diphone and suffix:
  smoothing in the range of one pitch period in a voiced environment, direct concatenation otherwise

- between syllables:
1. plosive meets homorgane plosive or nasal: first plosive is deleted, a pause of 100 ms is inserted
2. fricative meets voiced plosive: insertion of a pause of 50 ms
3. fricative meets homorgane fricative: first fricative is shortened by 30 ms
4. nasal, vowel, or liquid meets voiced plosive: insertion of a voiced pause of 50 ms length (one special unit for each place of articulation)
5. any sound meets unvoiced plosive: insertion of a pause of 60 ms (except when rule 1 applies)
6. vowel meets nasal, plosive, or fricative: the according diphone with the transition to the respective phoneme is selected
7. liquid meets fricative or plosive: a suffix with an according transition is chosen (plosive is cut off from the suffix)
8. nasal meets homorgane nasal or fricative: shortening by 30 ms
9. liquid meets itself: shortening by 50 ms
10. vowel meets accented vowel or vowel in next word: a pause of 25 ms is inserted to simulate the glottal stop
11. vowel meets unaccented vowel in same word: spectral interpolation over 80 ms is performed

Table 1 Concatenation rules currently implemented in the HADIFIX speech synthesis system.

duration phenomena modeled:
- phrase-final lengthening (depending on vowel type)
- phrase-initial shortening
- stressed syllable lengthening (depending on vowel type and stress level)
- vowel lengthening before /r/, vowel shortening before voiceless plosive
- adjusting syllable duration depending on number of phones in syllable
- shortening of polysyllabic words

intensity phenomena modeled:
- intrinsic intensity for vowels
- stressed syllable rising (depending on stress grade)
- intensity declination in an utterance

Table 2 Prosodic phenomena modeled by duration and intensity rules currently implemented in the HADIFIX speech synthesis system.

rules were put up heuristically. Table 2 shows the phenomena modeled by these rules.

F_0 contours are generated using Fujisaki's model [2,3]. A few standard parameters and some rules, which modify these parameters according to the context [3], are sufficient to produce an intonation contour that sounds quite natural.

All these rules are limited in scope to short utterances (two phrases max.). Prosody for longer utterances and whole paragraphs is currently determined by simple extrapolation of the situation for the short utterances.

5. EVALUATION - METHOD

The quality of the synthesized speech was evaluated under the aspects of intellegibility and naturalness [9]. Four "voices" were used in the evaluation: a female voice ("ANGEL"), a male voice ("STEFAN"), the female voice in a LPC-parametric representation ("LP-ANGEL"), and a human speaker ("MM").

Segmental intellegibility was determined using a rhyme test [10,11] with 40 items and 5 choices for each item (although one can surely question the usefulness of such a test for synthesis systems with units as large as DSs). The word level was investigated by phonetically balanced lists of 40 words. These words were embedded in a carrier phrase with the key word to be written down by the subjects. To measure text intelligibility, some items out of a collection of short, uniformly difficult texts [12] were presented for dictation. In all these tests, listeners were asked for their subjective rating as to how many words they thought they had not understood.

For the segment and word levels naturalness was simply accessed by global scalings on an 11-point scale (0-10). At the text level, a special test was designed using some other texts from the collection mentioned above [12], because far more factors influence naturalness at this level (especially prosody, but also the fluency of the concatenated speech). Subjective ratings over 10 attributes as semantic differentials were used as input for a principal component analysis with a varimax transformation of the two main components [13,14]. Table 3 lists the attribute pairs used in the test.

indistinct distinct	unnatural natural	unintelligible intelligible	unpleasant pleasant	hollow clear
slow fast	hesitating fluent	monotonous lively	flat plain	unmelodious melodious

Table 3 Attribute pairs used for the evaluation of the naturalness at the text level of the synthetic speech from the HADIFIX system.

6. EVALUATION - RESULTS

The outcome of the rhyme test (n=29) is shown in Figure 1. The differences between the voices are significant with respect to error rate ($F(3,112)=66.5$, $p<0.001$), naturalness ($F(3,112)=56.5$, $p<0.001$), and assumed error rate ($F(3,112)=25.5$, $p<0.001$); error rates as inverse measures of recognition rates were used for convenience of graphical display. A Tukey test confirmed that all differences between voices are significant ($p<0.05$), except those between ANGEL and STEFAN (both time-domain units).

Figure 2 displays the results of the word test (n=36). For reasons of time, the human voice was not included in this test. Here too, the differences are significant with respect to all parameters: error rate ($F(2,105)=240.3$, $p<0.001$), naturalness

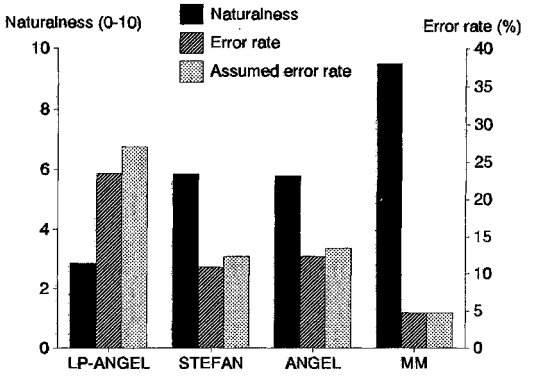

Fig. 1 Results of the rhyme test (n=29) (MM: human speaker; ANGEL and STEFAN: time-domain units; LP-ANGEL: LPC units).

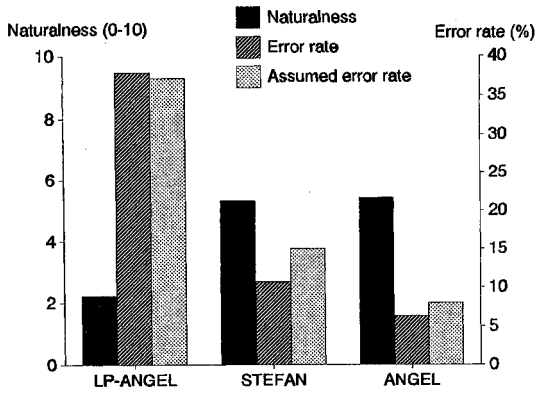

Fig. 2 Results of the word test (n=36) (ANGEL and STEFAN: time-domain units; LP-ANGEL: LPC units).

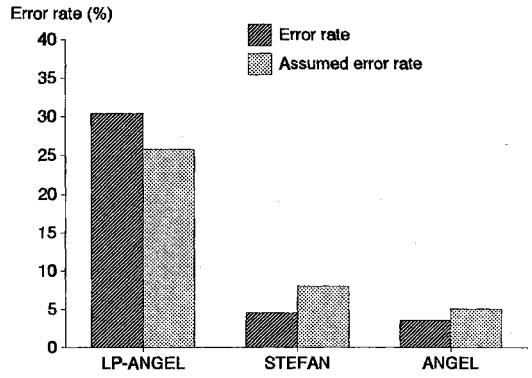

Fig. 3 Results of the text intelligibility test (n=23) (ANGEL and STEFAN: time-domain units; LP-ANGEL: LPC units).

(F(2,105)=51.2, p<0.001), and assumed error rate (F(2,105)=57.5, p<0.001). Again, the Tukey test showed significant differences (p<0.05) only between LP-ANGEL and the two time-domain unit voices.

In the text intelligibility test (Figure 3) (n=23), the human voice was again not involved for reasons of time. The differences for error rate (F(2,66)=81.0, p<0.001) and assumed error rate (F(2,66)=28.3, p<0.001) were significant only between LP-ANGEL and the two other voices ANGEL and STEFAN (Tukey test, p<0.05).

The test for naturalness in longer speech passages (Figure 4) (n=23) showed no such clear outcome, although significant differences exist for each attribute (F-test, p<0.05). Therefore a principal component analysis with a varimax transformation for the two most important components was performed. A reanalysis of the attribute ratings with transformation to the two components established significant differences only for the first component (F(3,88)=121.21, p<0.001), not for the second (F(3,88)=0.86, 0.5<p<0.25). One may interpret the first component as "quality" and the second as "individual preferences of the listeners". Thus, the first component may display the intended result of the test. Figure 5 shows the result of the transformation to the two components.

7. EVALUATION - DISCUSSION

One premier result is the clear superiority of time-domain units over LP-coded units. While the intelligibility of STEFAN and especially ANGEL is high (94 % on the word level, 97 % on the

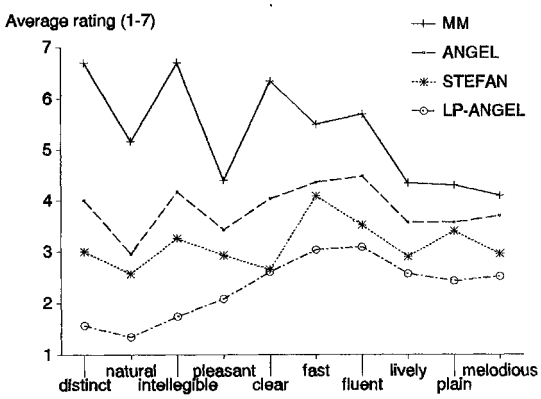

Fig. 4 Average ratings of the text naturalness test (n=23) (MM: human voice; ANGEL and STEFAN: time-domain units; LP-ANGEL: LPC units).

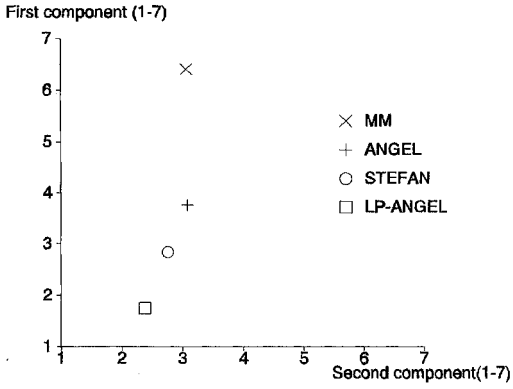

Fig. 5 Results of the text naturalness test transformed to the two principal components (MM: human voice; ANGEL and STEFAN: time-domain units; LP-ANGEL: LPC units).

text level), the differences in naturalness between the human voice and the synthetic voices are still large. Figures 6 and 7 comprise the overall results.

The error rate of LP-ANGEL increases severely between the closed-vocabulary rhyme test and the word test. The other voices seem good enough to compensate the higher difficulty of the task. The differences between the word and text levels are only gradual; context lowers error rate (as one could expect). The method of taking a dictation seems a good method to evaluate text intelligibility: it is handy, robust, and its result is not influenced by a different "world knowledge" of the subjects. It has a high ecological validity [12], compared with tests using isolated words or (even semantically anomalous) sentences.

A noticeable result is the good agreement between assumed and real error rates: the difference between these two error measures was found insignificant (t-test,p<0.05), except for one case, the word test for STEFAN.

The difference in naturalness between STEFAN and ANGEL is visible only at the text level. The ratings depend on the length of the tokens presented; shorter tokens caused better ratings than longer ones (STEFAN at text level is as bad as LP-ANGEL at the rhyme test), which is not suprising considering the larger number of possible error sources. Tests for naturalness must therefore inevitably contain longer parts of speech (especially for TTS systems, where syntax and semantics have to be reflected by

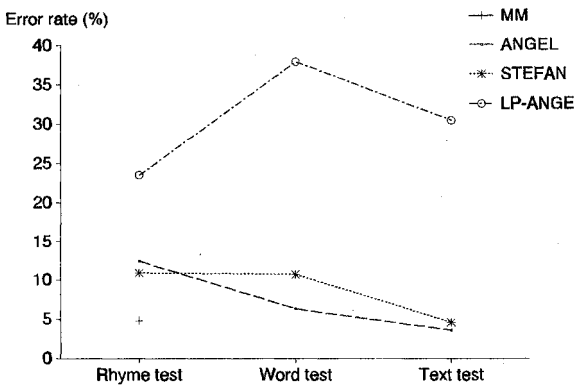

Fig. 6 Error rates of the voices in all tests (ANGEL and STEFAN: time-doiman units; LP-ANGEL: LPC units). MM (human voice) was not tested for word and text intelligibility.

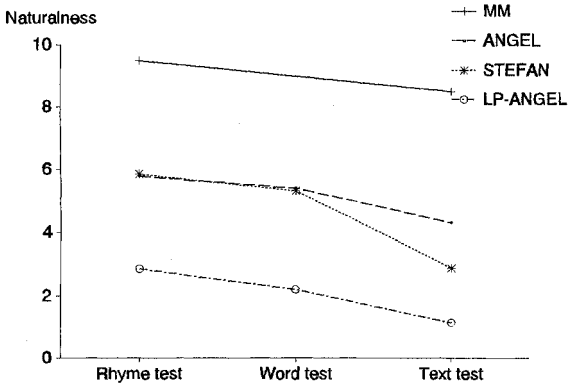

Fig. 7 Naturalness ratings of the voices (MM: human; ANGEL and STEFAN: time-d. units; LP-ANGEL: LPC units). For the text level the value of the first component is displayed after rescaling. MM was not tested at the word level.

the prosodic structure).

8. CONCLUSION

The evaluation proved the high quality of the speech generated by HADIFIX. Nevertheless, some improvements have to be made. Only a few points will be emphasized here:

- The current inventory has no special units for function words or reduced forms. Therefore, some words sound overarticulated. The simple cure is an addition of carefully chosen and recorded units representing the most common function words.
- Problems arise at unit boundaries in postvocalic liquids. A change of the splitting point in the final DS towards the voiced-unvoiced boundary is recommended.
- Prosodic rules must be extended to describe multi-phrasal utterances and text passages adequately.

For further points see the discussion in [4].

HADIFIX demonstrates in its current implementation that synthetic speech with nearly complete intelligibility and fair naturalness can be produced by concatenating time-domain syllable-oriented units of real speech. HADIFIX works in real time (except the spectral interpolation routine) on an ordinary PC-486 without any DSP boards. But one should remember that HADIFIX is not a text-to-speech system, but operates on a phonemic input with accent markers.

ACKNOWLEDGEMENTS

The authors thank Dieter Stock for his fine SONA program, also Bernd Möbius and Matthias Pätzold for their intonation rules and parameters. This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG).

REFERENCES

[1] Moulines, Eric / Charpentier, Francis (1990): "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones." Speech Commun. *9*, 453-467

[2] Fujisaki, Hiroya (1988): "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour." In *Vocal physiology: voice production, mechanisms and functions (= Vocal fold physiology, Vol.2)*; ed. by Osamu Fujimura, 347-355 (Raven Press, New York)

[3] Möbius, Bernd / Pätzold, Matthias (1992): "F_0 synthesis based on a quantitative model of German intonation", this volume.

[4] Hess, Wolfgang (1992): "Speech synthesis - a solved problem?". Invited paper, presented at EUSIPCO-92 (Brussels)

[5] Fujimura, Osamu (1976): Syllable as the unit of speech synthesis (Intl. Rept., AT&T Bell Labs, Murray Hill, NJ, USA)

[6] Dettweiler, Helmut / Hess, Wolfgang (1985): "Concatenation rules for demisyllable speech synthesis." Acustica *57*, 268-283

[7] Portele, Thomas / Steffan, Birgit / Preuß, Rainer / Hess, Wolfgang (1991): "German speech synthesis by concatenation of non-parametric units." In *Proc. EUROSPEECH-91*, 317-320 (Genova, Italy)

[8] Portele, Thomas / Sendlmeier, Walter F. / Hess, Wolfgang (1990): "HADIFIX: a system for German speech synthesis based on demisyllables, diphones, and suffixes." In *Proc. of the ESCA Workshop on Speech Synthesis*, 161-164 (Autrans, France)

[9] Portele, Thomas / Sendlmeier, Walter F. / Stock, Dieter / Hess, Wolfgang / Steffan, Birgit / Preuß, Rainer (1992): "Evaluierung des Sprachsynthesesystems HADIFIX." To appear in *Fortschritte der Akustik: DAGA-92* (Berlin, Germany)

[10] Sotscheck, Jochem (1982): "Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachübertragungsgüte." Der Fernmeldeingenieur *36* (4/5), 1-45

[11] Sendlmeier, Walter F. / von Wedel, Hasso (1986): "Ein Verfahren zur Messung von Fehlleistungen beim Sprachverstehen - Überlegungen und erste Ergebnisse." Sprache, Stimme, Gehör *10*, 164-169

[12] Sendlmeier, Walter F. / Holzmann, Ulrike (1991): "Sprachgütebeurteilung mit Passagen fließender Rede." In *Fortschritte der Akustik: DAGA-91*, 969-972 (Bochum, Germany)

[13] van Bezooijen, Renée / Pols, Louis C. W. (1991): "Performance of text-to-speech conversion for Dutch: A comparative evaluation of allophone and diphone based synthesis at the level of the segment, the word, and the paragraph." In *Proc. EUROSPEECH-91*, 871-874 (Genova, Italy)

[14] Bappert, Veronika / Ehlert J. (1991): "Sprachgütebeurteilung mit Hilfe des Semantischen Differentials." In *Fortschritte der Akustik: DAGA-91*, 1105-1108 (Bochum, Germany)

[15] Stock, Dieter: personal communication