TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Development and perceptual evaluation of a timing module for German diphone speech

Citation for published version (APA):
Rump, H. H. (1991). *Development and perceptual evaluation of a timing module for German diphone speech*. (IPO rapport; Vol. 829). Instituut voor Perceptie Onderzoek (IPO).

Document status and date:
Published: 02/10/1991

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 05. Oct. 2023

Institute for Perception Research
PO Box 513, 5600 MB Eindhoven

HHR/hhr 91/02
02.10.1991

Rapport no. 829

Development and perceptual evaluation of a timing module for German diphone speech

H.H. Rump

Development and perceptual evaluation of a timing module for German diphone speech

Willem Rump

september 1991

Abstract

In this paper a description will be given of the development and perceptual evaluation of a timing module for the duration control of synthesized German diphone speech. For the perceptual evaluation a number of sentences was synthesized in three timing versions: 1. with natural timing structure, 2. with duration control by rule and 3. without nearly any duration control. The sentences were synthesized using male voice diphones and female voice diphones. Listeners had to judge the acceptability of the sentence timing, scoring on a 10 point scale. The sentences were presented separately. The sentences having the natural timing structure were supposed to score highest. The hypothesis was that sentences having a rule generated timing structure would score as high as the ones having a natural timing structure.

The results showed no significant difference between the three timing versions. There was a large effect of the sentences tested: concerning the ten sentences that were synthesized with the male voice diphones, there was a slight tendency for the natural timing version scoring higher than the other two versions. However, looking at the data for the five sentences tested with the female voice diphones, there was a strong tendency for the rule based timing being most acceptable. This also held for the five sentences when synthesized with the male voice diphones.

The conclusion is that the corpus of tested sentences should be much larger to get more reliable results. Due to several other factors, including the overall quality of the synthetic speech used, and the excessive shortness of the sentences tested, the results were less clear than expected.

(See next page for Dutch and German versions of the abstract.)

Samenvatting

In dit verslag zal een beschrijving gegeven worden van de ontwikkeling en de perceptieve evaluatie van een duurmodule voor het beregelen van de timing van Duitse difoonspraak. Voor de perceptieve evaluatie werd een aantal zinnen gesynthetiseerd in drie timingversies, 1. opgelijnd aan natuurlijke spraak, 2. met timing volgens regels en 3. nagenoeg zonder duurberegeling. De zinnen waren eenmaal met de Huber-difonen en eenmaal met de Ursula-difonen gesynthetiseerd. De proefpersonen moesten op een tienpuntsschaal aangeven hoe goed zij elke timingversie vonden. De zinnen werden apart aangeboden. Aanname was, dat de versie met de natuurlijke timing favoriet zou zijn. De hypothese was dat zinnen die met de timingregels gesynthetiseerd waren even hoog zouden scoren als zinnen met de natuurlijke timing.

Het resultaat van de evaluatie was, dat er geen significant verschil was in de beoordeling van de drie timingversies. Er ging veel invloed uit van de zinnen die getest werden: de data voor de tien Huber-zinnen lieten de tendens zien, dat de versie met de natuurlijke timing hoger scoorde dan de beide andere versies. De data van de vijf Ursula-zinnen laten echter een sterke tendens zien dat de timing volgens regels het meest acceptabel gevonden werd. De data voor deze vijf zinnen met Huber-difonen bleken diezelfde tendens te vertonen.

De konklusie is, dat het corpus van geteste zinnen uitgebreider zou moeten zijn om een betrouwbaarder resultaat te verkrijgen. Daarnaast speelden andere faktoren een rol, zoals de matige kwaliteit van de gebruikte gesynthetiseerde spraak, en het gebruik van tamelijk korte zinnen, zodat de resultaten veel minder duidelijk waren dan verwacht.

Zusammenfassung

In diesem Bericht wird die Entwicklung und die perzeptive Evalierung eines Dauersteuerungsmoduls für Deutsche Diphonsprache beschrieben. Einige Sätze wurden synthetisiert in drei Dauerversionen: 1. mit natürlichen Dauerverhältnissen, 2. mit nach Regeln generiertem Timing, und 3. ohne nennenswerte Dauersteuerung. Die Sätze wurden einmal mit Huber-Diphonen (Männerstimme) und einmal mit Ursula-Diphonen synthetisiert. Versuchspersonen wurden gebeten, auf einer 10-Punkte-Skala anzudeuten, wie akzeptabel sie jede der einzelnen Timingversionen fanden. Die Sätze wurden ihnen einzeln vorgespielt. Zu erwarten war, daß die Version mit dem natürlichem Timing am meisten bevorzugt werden würde. Die Hypothese war, daß die Sätze, die mit dem Timing nach Dauerregeln synthetisiert worden waren, gleich hoch bewertet werden würden wie die Sätze mit dem natürlichem Timing.

Die Ergebnisse zeigten, daß es keinen signifikanten Unterschied zwischen den Bewertungen der drei Timingversionen gab. Der Einfluß der einzelnen Sätze war besonders groß : für die zehn Sätze, die mit den Huber-Diphonen synthetisiert waren, gab es eine leichte Tendenz, daß die Version mit dem natürlichem Timing von den Versuchspersonen bevorzugt wurde. Wenn man aber die Daten der fünf Sätze, die mit den Ursula-Diphonen synthetisiert waren, besieht, zeigt sich eine starke Tendenz, daß die Version mit dem Regelgesteuertem Timing bevorzugt wurde. Dasselbe galt für diese Sätze, die mit den Huber-Diphonen synthetisiert worden waren.

Die Schlußfolgerung ist, daß mit einer größ eren Zahl von Sätzen ein deutlicheres Ergebnis zu erziehlen wäre. Daneben haben unseres Erachtens die mäßige Sprachkwalität der LPC-synthetisierten Sätze, und die Verwendung sehr kurzer Sätze dazu geführt, daß die Ergebnisse weniger eindeutig waren als erwartet.

Contents

1 Introduction

1.1 Timing in speech

In the present paper we will describe the development and perceptual evaluation of a timing module for the duration control of synthesized German diphone speech. Duration control is very important for the perceived naturalness of synthetic (computer) speech. If speech is too fast it will be hardly understandable, and if it is too slow it will be very boring. The overall speech rate is perceived as tempo.
There are also sentence internal variations in speech rate caused by pauses and lengthening of speech segments. Their occurrence depends on speaker's intentions and on linguistic structure. Variations in speech rate are perceived as rhythmic pattern of sentences.

In order to make synthetic speech sound more natural we developed a module for the manipulation of sentence timing. The timing module contained duration rules extracted from natural speech. The duration of synthetic speech segments was manipulated according to these rules.
In the following sections we will give a short review of a method used to discover timing regularities in naturally spoken sentences. Subsequently, we will sum up the rules for the duration control and outline the structure of a newly developed timing module. Finally, we will describe the perceptual evaluation of the naturalness of speech synthesized using the timing module, and we will discuss the results of the experiment in which three timing versions of each sentence were compared. Speech synthesized with 'natural' durations was supposed to be most acceptable, and speech nearly without timing manipulation was supposed to be least acceptable. The hypothesis was, that speech with duration control by rule would be as acceptable as speech with a timing structure that was derived from natural speech.

1.2 Speech synthesis by means of diphones

At IPO synthetic speech is generated by means of diphones. Diphones are segments of natural speech in which the last part of a phoneme, the transition and the first part of the second phoneme are preserved. They are stored on computer disk in LPC-coded form and they are concatenated and synthesized in order to produce synthetic speech. The boundary between the two phoneme parts making up a diphone is marked in order to enable measurement of phoneme durations.
Two kinds of diphones are available: diphones segmented from a male voice (Huber) and diphones segmented from a female voice (Ursula). Both kinds of diphones are standardized. Standardization of diphones means

that phoneme parts of a given phoneme have the same duration in all diphones containing that phoneme in a particular position. For example, the /a/-part in the /ma/-diphone had the same duration as the /a/-part in all other C-/a/- or V-/a/-diphones.
After concatenation of the diphones the resulting sentences (parameter files) are provided with an F_0-contour. Finally, the sentences are synthesized in order to produce speech output.

2 Timing stylization

The timing problem in speech is very complex. Duration of segments depends on many factors, some causing lengthening, others causing shortening (see for example van Coile, 1990, and van Santen, 1990). These factors also seem to interact and their perceptual relevance is not always clear.
Therefore, we developed the method of timing stylization, in which we abstract as far as possible from segmental factors, in order to find suprasegmental factors that influence segmental durations. Those factors were to be discovered in natural spoken sentences, since their timing was supposed to be optimally natural.
Regularities in the timing of naturally spoken sentences were found using the method of 'Close Copy' stylization, analogous to the intonation research method developed at IPO (De Pijper, 1983). The basic idea was that there is an inherent or mean duration of each segment, which is mainly influenced by suprasegmental phenomena.
For each phoneme the mean duration was found from the segment's duration in natural speech (see section 2.2). Those mean values were listed in a duration table (like the so-called 'Klatt-table' in Appendix A).
Regularities in sentence timing describe the circumstances in which the duration of a segment is different from its mean duration. In the next section we will describe a method to represent sentence timing.

2.1 Sentence timing

The timing structure of a sentence is described at the segmental level, i.e. by the durations of phonemes and pauses in the sentence. At IPO programs are available to display graphically the timing structure of a sentence (see Figure 1).
The durations of segments are graphically represented by label values. A label value determines the frame duration (in ms) that will be considered

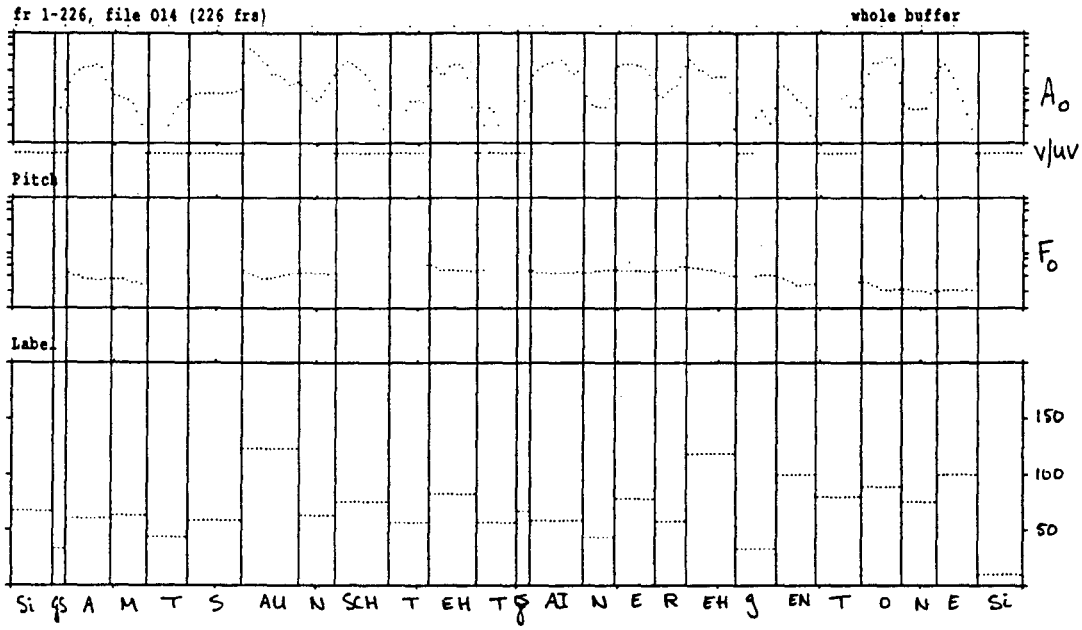

Figure 1: *Timing structure of a diphone sentence (vertically phoneme boundaries are indicated): S64: Am Zaun steht eine Regentonne, "At the fence there is a water-butt". The horizontal lines represent label values.*

during synthesis of a speech segment[1]. A label value of 50 % means that the segment will have half the duration of a formerly specified duration (i.e. the duration of the segment after 'raw' diphone concatenation, or of the value listed in a duration table). A high label value indicates that the duration of the synthesized segment will be long, a low value that it will be short, relative to a specified duration.

2.2 Timing alignment and Close Copy stylization

We wanted to implement rules for durational control of synthesized speech segments. In order to find perceptually relevant factors influencing segment durations in natural speech, the timing structure of natural spoken sentences was measured. (The factors might be segmental and suprasegmental as well.) Measuring the timing structure was enabled by transferring the timing structure of a natural spoken sentence onto the matching diphone concatenated sentence. This method was called 'timing alignment' ('oplijnen'): the segments in the diphone sentence got exactly the same durations as the corresponding segments in the original spoken sentence. The segmental durations could than be stored or printed.
In an earlier experiment we used the timing alignment method in order to determine the mean durations of speech segments in natural spoken

[1] A frame is the time period during which the speech parameters are not renewed: for example, a frame duration of 10 ms means new parameters every 10 ms.

sentences. The values we found were stored in a so-called Klatt-table (see Appendix A).

After concatenating diphones the resulting phoneme durations can graphically be displayed as label values. From the previous experiment we concluded that it is possible to draw a timing contour connecting the labels (stylization) without introducing perceptual differences in the timing of the sentence. A timing contour that was built of as few straight lines as possible connecting the original labels and that resulted in a perceptual identical sentence timing, was called a 'Close Copy' of the original timing contour. (For a detailed description of the timing alignment and stylization processes see Adriaens-Porzig and Rump, 1990).

2.3 Durational regularities

Factors influencing durations of speech segments were to be found in naturally spoken sentences. For German the corpus of the hundred so-called 'Sotschek'-sentences was available. They were read by a German newsreader. The sentences were manually segmented and provided with a phonemic transcription in order to make them suitable for timing alignment.
After timing alignment we found durational regularities by Close Copy stylization of fourty sentences from the Sotschek-corpus.

Mean level

After having stylized the fourty Sotschek-sentences we found a mean level of phoneme duration at label value 55 or 60 (i.e 55 or 60 % of the duration values specified in the Klatt-table. In general, the first part of a sentence (being about the first fifteen or twenty phonemes) was spoken a little faster than its remainder. From this level only lengthening of phonemes took place. However, after having synthesized some sentences using this mean level of phoneme duration, the resulting speech sounded much too fast. After raising the mean level to about 70 or 80 percent, the sentences sounded much better. But then there were parts of the sentences that were spoken far too slow.
Thus, we had to look for a common feature of the words that should be shortened. The best-fitting group of words was the group of so-called 'function words'. Therefore, a shortening rule for function words was introduced. The group of function words will be discussed at the end of this section.
Another consequence of raising the mean level was that the amount of lengthening of phonemes had to be reduced in order to avoid syllables sounding unnaturally long.

Lengthening at sentence accent

In case of a sentence accent, (part of) the accented syllable was lengthened. The amount of lengthening depended on the phonological length of the syllable nucleus: 'long' vowels would be lengthened more than 'short' vowels (a schwa never will be in an accented syllable). A short test showed that lengthening the nucleus of a syllable was nearly enough to reach the right duration perception for the whole syllable. Only the consonant preceding the syllable nucleus had to be lengthened a little in case of an accented syllable. If a short accented vowel was followed by a voiced consonant, that consonant should also be lengthened a little.
A remarkable phenomenon was the case of a schwa following an accented syllable: it was lengthened if there was only one consonant following the accented syllable nucleus, for example: 'Regentonne ("waterbutt"); if there were more consonants, the schwa would not be lengthened, for example: 'Ärzte ("doctors").
In case of sentence accent on the last vowel in the sentence, there would be some extra lengthening of the last phonemes.
Because there were no rules available to assign sentence accents automatically, the accents had to be marked in the input sentence. In case of grapheme input this had to be done by putting a quote before the accented word. In case of phoneme input the accentuation marker (') had to be put before the nucleus of the accented syllable (e.g. 'Vater versus F'AHTER).

Prepausal and sentence-final lengthening

Because there were no compound sentences in the Sotschek-corpus, no rules were found for prepausal lengthening. We would suggest to apply the rules for sentence final lengthening in case of a pause after a comma. Different situations were possible before pauses (indicated by a comma) or at the end of sentences, depending on the kind of phonemes in those positions. We will call prepausal or sentence final phonemes 'final' if they are the very last phoneme before the boundary, otherwise we will call them 'last' phonemes. Possible sequences were:

- -V (= final non-reduced vowel),
- -V(C)C (= last vowel and final consonant),
- -V(C)Schwa (= last vowel and final schwa),
- -V(C)Schwa(C)C (= last vowel, last schwa and final consonant).

For prepausal or sentence-final lengthening we found the following regularities:

- A sentence-final non-reduced vowel or schwa will always be lengthened.
- A sentence-final consonant will also be lengthened, and a strident consonant will be lengthend more than a non-strident consonant,
- If the last vowel in the sentence is non-final it will be lengthend less than a sentence-final vowel,
- If the last vowel is a schwa, the last non-reduced vowel preceding it will also be lengthened,
- Duration of optional consonants, i.e. (C), will not be affected.

The extent of the lengthening of a vowel depended on its phonological length: a phonologically long vowel was lengthened more than a phonologically short vowel, and a short vowel more than a schwa.

Function words

Some words did not receive any prominence usually. These words were listed together in a separate word category of 'function words'. They included: articles, pronouns, prepositions, interjections, auxiliaries and copulas. This separate word category was also introduced in the duration analysis of Dutch and Japanese sentences (van Coile, 1990, Kaiki, Takeda and Sagisaka, 1990), although the perceptual relevance of a special duration rule for it was not tested. (The category of function words was in those cases introduced on statistical grounds.)
In some cases function words had sentence accent (like: 'Wer möchte 'keinen Kuchen?, "Who would not want cake?"). In that particular case they would be lengthened instead of shortened.
The way to mark a word as a function word was very ad hoc: a backslash ('\') was put in the phoneme input before the nucleus of the first syllable of the function word.

Question sentences

In case of a question the lengthening of the last phonemes of the sentence would take place as if there were an accent on the last vowel of the sentence: there was some extra lengthening of the final consonant, of the last or final vowel and of the last or final following schwa as well.

Thus, the factors causing a perceptually relevant lengthening or shortening relative to the mean duration listed in the Klatt-table, were:

- Position of the phoneme in the sentence (in the first part of a sentence the mean segment duration will be lower than in the second part of a sentence),

- Position of the phoneme relative to a sentence boundary (pause or end) (in the last word before a comma or before the end of a sentence some phonemes will be lengthened additionnally),
- Presence of an accent-lending pitch movement (a sentence accent will cause lengthening of some phonemes in the accented syllable),
- Presence of vowel-schwa clusters[2] (the vowel will be shortened),
- Presence of plosive clusters (the first plosive will be shortened),
- Word class (phonemes in function words will be shortened),
- Kind of sentence (a question sentence will cause additional final lengthening).

The some factors turned out to be on the suprasegmental level, while the others turned out to be on the segmental level. Together they were the factors included in the module for durational control of German diphone speech. In the next section we will describe the newly developed timing module.

3 The timing module

3.1 Structure of the module

Durational regularities were found by stylization of sentence duration. The duration rules will act on the segmental level.
The newly developed timing module was written in PASCAL and it was part of the diphone speech synthesis system DS. The position of a timing module in DS is after concatenation of the diphones and before speech synthesis: the sentences then pass through a timing module and through an intonation module.
The following steps were included in the timing module: first, a table specifying phoneme durations and a table containing phoneme features will be read. Subsequently, features like 'last_vowel', 'last_schwa' and 'last_consonant' will be assigned and, finally, the duration rules will apply. The steps will be discussed separately in the following sections.

3.1.1 Phoneme durations

The timing module was set up in order to change the label values of the phonemes before the sentence was synthesized. The label values were specified by a Klatt-table, in which absolute phoneme durations were given in

[2] In the present paper we will call the syllabic vowels (-EL, -EM, -EN and -ER) together with the reduced vowel (-E): 'schwa'.

milliseconds (see Appendix A). These durations were the values to be manipulated by the newly developed duration module. (In another timing module phoneme durations were determined, e.g. by the way that diphones were segmented from natural speech.) The specific values in the Klatt-table were chosen so that it was possible to draw a very simple Close Copy timing contour in any timing aligned sentence (see for example Figure 2).

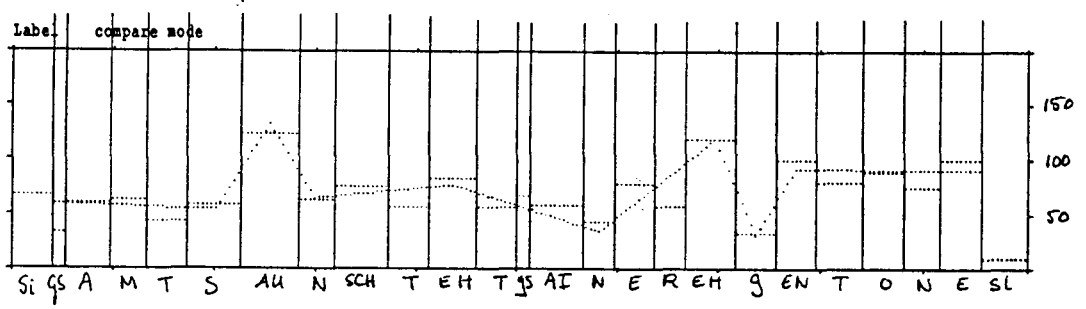

Figure 2: *A timing aligned sentence ('natural' durations) and a stylized timing contour for the same sentence (vertically phoneme boundaries are drawn). (S64: Am Zaun steht eine Regentonne, "At the fence there is a water-butt".)*

3.1.2 Phoneme features

Phoneme features were read from a phonetic feature table (see Appendix B). Since standardized diphones were used, it became possible to distinguish different phonological phoneme classes, e.g. plosives, long vowels etc., to be treated the same. Standardization was namely realized by fixing the number of frames of the phoneme parts in diphones, so that phonemes within one class got the same number of frames (Vogten, 1988). Before any durational rule would apply, the features 'last_consonant', 'last_schwa' and 'last_vowel' had to be assigned to the appropriate phonemes; in case of a question sentence the feature '+ sentence_accent' had to be assigned to the last non-reduced vowel in the sentence.

3.2 Implemented rules

In this section we will sum up the rules implemented in the duration control module for the speech synthesis program DS. We found the regularities in sentence-internal duration differences using the timing stylization method and afterwards we formalized the regularities as duration rules. The following rules applied to the phonemes in a sentence:

- Assignment of a mean (frame) duration for the first twenty phonemes of the sentence (except for sentences shorter than 20 phonemes),
- Assignment of a mean (frame) duration for the remaining phonemes in the sentence (or for phonemes in the short sentences),
- Lengthening of accented long vowels,
- Lengthening of accented short vowels,
- Lengthening of prepausal or sentence final long vowels,
- Lengthening of prepausal or sentence final short vowels,
- Lengthening of prepausal or sentence final schwa,
- Shortening in vowel-schwa clusters,
- Shortening in plosive clusters,
- Shortening of function words,
- Lengthening in question sentences.

Most of the rules cited above could be manipulated interactively during the use of the speech synthesis program DS. Shortening of phonemes in the particular clusters was also done automatically by the program.
In case of lengthening of vowels also the duration of other segments that were closely related to the lengthened vowel had to be manipulated. For example, in the case of lengthening of a short accented vowel the following voiced consonant had to be lengthened too. The appropriate lengthening of the so-called 'related' phonemes was executed automatically by the program.
In the next section we will discuss a perception experiment that we performed in order to perceptually evaluate the speech quality of sentences that had a timing structure generated by the newly developed timing module.

4 Perceptual evaluation

4.1 Method part I: Male voice

4.1.1 Material

In order to perceptually evaluate the newly developed timing module we chose ten sentences from the Sotschek-corpus (Appendix C). The sentences were synthesized with the standardized Huber-diphones using the phonemic transcription of the original spoken sentences. Three versions of each sentence were synthesized, each version having a different internal timing:

- Version 1: sentences with the timing structure of the naturally spoken sentences (timing aligned), hereafter: NAT,
- Version 2: sentences with the timing structure assigned by the newly developed timing module, hereafter: NEW,
- Version 3: sentences with the phoneme durations set to 80 % of the 'raw' phoneme durations (i.e. of the phoneme durations after simply concatenating the diphones), hereafter: RAW.

The overall duration of the sentences did not differ much for the three versions (Appendix D). In general the durations of the NAT versions were the shortest and the durations of the RAW versions were the longest. Overall, the durations lay between 1.2 and 2.5 seconds.

The parameters of the duration module used to generate the NEW version were:

- duration of the phonemes in the first part of the sentence: 70 % of the table values of phoneme durations,
- second part of the sentence: 80 % of the table values,
- accented long vowels: mean value + 20 % of the table value,
- accented short vowels: mean value + 20 %,
- last long vowel: mean value + 20 %,
- last short vowel: mean value + 20 %,
- last schwa: mean value + 20 %,
- function words: mean value - 30 %.

The extra lengthening of 'related' phonemes and of phonemes under particular circumstances, like, for example, accentedness of the last vowel, was executed by the program using fixed factors. An example of a sentence with the timing structure after having passed the duration module is shown in Figure 3.

The intonation of the sentences was generated automatically using German intonation rules. Differences in intonation originating from timing alignment were corrected manually during visual inspection to assure identical intonation over the three timing versions.

4.1.2 Subjects

The subjects that took part in the perception experiment were German native speakers. The test took place at the Institute for Communication and Phonetics (IKP) in Bonn, Germany. The subjects were 17 employees and students of IKP, and some of them were experienced in listening to

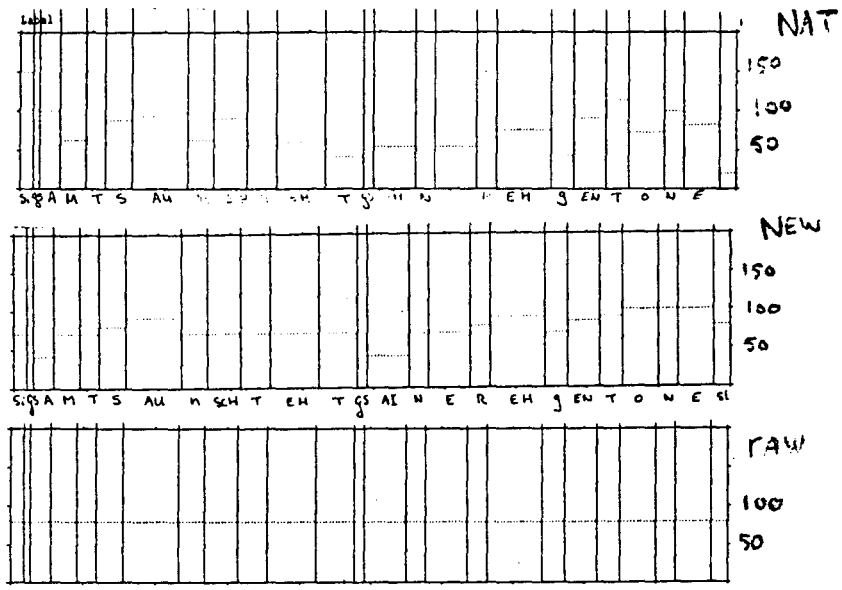

Figure 3: *Segment labels for the three versions (*NAT, NEW *and* RAW, *respectively) of a sentence synthesized with Huber diphones (vertically phoneme boundaries are drawn).*

synthetic speech.
In order to discard data of subjects that proved to be quite unstable in their judgements the following criterion was chosen: if the mean standard deviation of the scores was larger than 1.7, the data of the subject were excluded.

4.1.3 Procedure

In order to test the naturalness of the sentence timing of the three differently manipulated timing versions we performed a scaling experiment. The subjects were instructed to listen to separately presented sentences and to judge how natural tempo and rhythm of the speech sounded to them. They had to indicate how acceptable the timing of each sentence was on a scoring form. The scores ranged on a scale from 1 to 10, 1 being the lowest score and 10 the highest.
We decided to present the test items separately in order to avoid a temporal order effect like the one found in a previous (pairwise comparison) experiment. In the previous experiment it was shown that subjects could hear very small differences in timing between two timing versions.
Each of the ten sentences was synthesized in three versions, making up 30 items. We randomized twice a block of thirty sentences and then repeated this sequence backwards, which resulted in 120 test items. Each version of a particular sentence was thus included four times in the experiment.

(From experiments with visual stimuli it was known that scaling experiments with at least four repetitions, i.e. four judgements per stimulus, would give reliable results.) Separate presentation of the stimuli (instead of pairwise) would also reduce the amount of time necessary to counter balance for order (so that some time was left in which we could perform a second experiment which will be described in section 4.4.)
Between two successive items there was a pause of four seconds, in which the subjects had to make their decission and to write their answer down. After every five items a warning tone was included serving as an orientation point for the subjects. The total duration of the test was approximately 18 minutes.
Before starting the test the subjects listened to twenty introductory items (all of them also appeared in the test) to get used to the diphone speech quality and to learn how to use the scoring scale.
In a previous perception experiment the subjects complained that without a break the test was too long. Therefore, we introduced a short pause halfway in the present test (after the subjects had heard 60 test items).

4.2 Results male voice

The total amount of scores was 17 (subjects) x 10 (sentences) x 3 (versions) x 4 (repetitions) = 2040. First, we inspected the raw data in order to get an impression how certain subjects had been in assigning scale values to the stimuli. It turned out that only three subjects had a mean standard deviation larger than 1.7 . Their data were discarded so that 14 subjects remained. The total sum of scores was then 1680.
Further inspection of the raw data was done in order to see how much the subjects liked the diphone speech and how they judged the differences in sentence timing. Subsequently, we converted the raw scores into scale values, taking together the scores of the whole group of subjects.
In the statistical analyses of the data the limit to reach significance was set to 5 percent.

4.2.1 Raw data

The subjects judged the acceptability of the timing of the sentences on a scale ranging from 1 to 10. Score 10 was the highest, score 1 the lowest. The total mean score was 4.84. The means and standard errors of the mean (SEM's) for the three versions are in table I. The mean scores for the ten sentences divided over the three version are listed in appendix E. Only one sentence was judged to be very good: S64 ('Am Zaun steht eine Regentonne', "At the fence there is a water-butt", mean score 6.9) and just one sentence was judged to be sufficiently good: S77 ('Der Bahnhof liegt

Table I: *Mean scores and SEM's for the three versions (14 subjects, 10 sentences, N=140).*

Version	mean score	SEM
NAT	5.01	.07
NEW	4.77	.08
RAW	4.75	.07

sieben Minuten entfernt', "The railway station is seven minutes away", mean score 5.8, only version NAT being judged to be less good: mean score 4.9).
The mean scores of all the other sentences were between 3.6 and 5.2, which indicates that most of the sentences were judged below the mean of the score scale.
For the purpose of analyzing the scores with the statistically very powerfull analysis of variance the scores were converted into scale values.

4.2.2 Scale values

The conversion of the raw data into scale values was done using the program SCALES2. The program converted the scores on interval level into scale values on ratio level (Edwards, 1957). Scale values were obtained by converting the scores of all items into a psychological continuum scale. The scale value of each item was determined on the basis of the scores of all subjects for that item. The between-subjects variance was thus eliminated.
After the conversion the following number of scale values remained: 10 x 3 x 4 = 120. We applied an analysis of variance (ANOVA) to the scale values using the statistical package ALICE. The analysis was repeated twice, each time with another relevant replication factor: Sentences and Repetitions. The factor Sentences was chosen in order to allow generalization over more than the ten sentences included in the experiment. The factor Repetitions was chosen since within-cell variance was expected to be smallest over repetitions, so that most significant effects would result. First, we will discuss the outcome of each ANOVA seperately. Subsequently, we will discuss them together.

Replication factor: Sentences (10)

The three effects tested were: two main effects (Versions and Repetitions) and their interaction effect. The Version effect turned out not to be significant at all. The Repetition effect turned out to be highly significant

($F_{(3,27)} = 10.91, p < .01$). As can be seen in Figure 4 the mean score were much higher during the first two repetitions than during the latter two.

The interaction effect was not significant, what meant that the Repetition effect was in the same direction for all versions: the later in the test, the lower the scores.

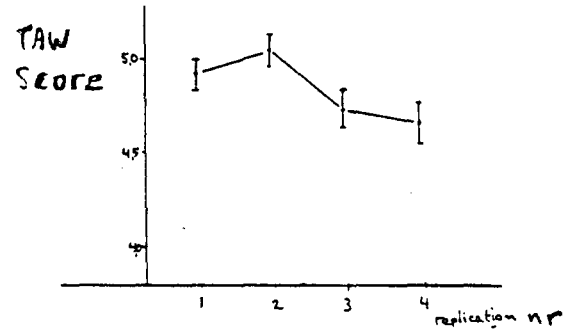

Figure 4: *Mean scores and SEM's per repetition for the male voice (14 subjects, 10 sentences, 3 versions, N = 420).*

Replication factor: Repetitions (4)

The three tested effects were: Version effect, Sentence effect, Sentence x Version effect. The effect of Versions turned out to be significant ($F_{(2,6)} = 6.69, p = .03$). As can be seen from Table I the mean score for the NAT version was significantly higher than for both of the other versions.

The effect of Sentence turned out to be highly significant ($F_{(9,27)} = 56.7, p < .01$). Some sentences got sinificantly higher scores than other sentences.

The effect of the interaction between versions and sentences was also highly significant ($F_{(18,54)} = 5.16, p < .01$). It indicated that in different sentences different timing versions resulted in the highest scores. A post-hoc comparison was performed in order to analyze per sentence what version got the highest scores.

4.2.3 Post-hoc analysis

In order to find out for what versions of which sentences the Version effect was significant we applied the Scheffé method of post-hoc comparisons to the data. First, a yard stick was calculated and then the differences between the version means per sentence were compared to this yard stick. The following significant differences in scale values were found:

- S26 'Im Geschäft stehen viele Leute' ("In the shop there are many people"): NAT better than RAW,

- S38 'Die Ärtze sind damit gar nicht einverstanden' ("The doctors don't agree with that at all"): NAT better than NEW and NAT better than RAW,
- S77 'Der Bahnhof liegt sieben Minuten entfernt' ("The railway station is seven minutes away"): NEW better than NAT and RAW better than NAT.

4.3 Discussion of part I

The results showed that all but one (S64 'Am Zaun steht eine Regentonne', "At the fence there is a water-butt") of the sentences got scores below 6. There was a clear overall effect of sentences. Some sentences got higher scores for the three versions than other sentences. Two sentences were judged to be extremely bad:
- S100 ('Das war jetzt aber ein schöner Tag', "That was a lovely day now", mean score 3.6), and
- S71 ('Rückt die Stühle an den Tisch', "Pull the chairs to the table", mean score 3.8).
An auditory inspection of these two sentences showed that S100 had too much function words close together, making the sentence too fast, and that in S71 the plosive cluster /KTD/ in 'Rückt die' was spoken too slow in two of the three versions.

The Repetition effect was also clear: the longer the people listened to the diphone speech, the lower the acceptability was judged. This result was against the expectation: generally, subjects get used to diphone speech so that they will judge it more acceptable.

For the versions there was no clear overall effect. The differences for sentences S26 ('Im Geschäft stehen viele Leute', "In the shop there are many people") and S38 ('Die Ärtze sind damit gar nicht einverstanden', "The doctors don't agree with that at all") were as expected: the version with the timing of the naturally spoken sentences was getting the highest scores. But for the other eight sentences there was no (significant) difference between the versions, or worse: in case of sentence S77 ('Der Bahnhof liegt sieben Minuten entfernt', "The railway station is seven minutes away") the version with natural timing was getting scores that were significantly lowest. An auditory inspection of that sentence revealed that the last two words came too close after each other.
In order to see whether the kind of diphones used in the experiment influenced the outcome of the test we performed a second experiment. It will be described next.

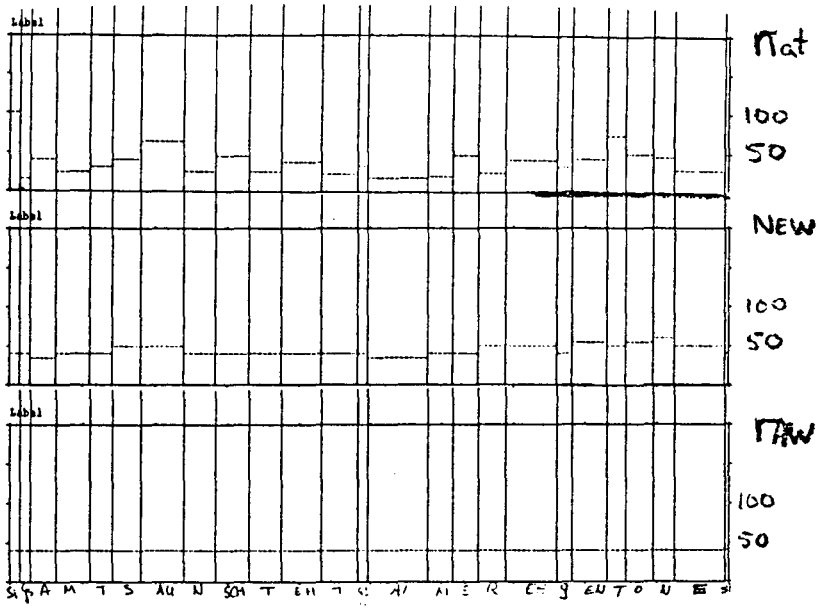

Figure 5: *Segment labels for the three versions (*NAT, NEW *and* RAW, *respectively) of a sentence synthesized with Ursula diphones (vertically phoneme boundaries are drawn).*

4.4 Method part II: Female voice

At IPO a new set of diphones was generated segmented from the recordings of a female voice (Ursula Adriaens-Porzig). These diphones also had standardized phoneme durations to enable application of straightforward methods of timing manipulation. An important difference between the two kinds of diphones was that female voice diphones were much longer than male voice ones. Therefore, the female voice diphones had to be much more compressed than the male voice diphones to reach the same rate of the synthesized speech.
Because we were interested in the general applicability of the timing rules, a set of female voice sentences was synthesized for a perceptual evaluation. The results of this test will be compared to the results of part I (the Huber, i.e. male voice, diphones).

4.4.1 Test sentences, subjects and procedure

The test sentences for the second part of the present perception experiment were a subset of the sentences used in part I: the numbers are S16, S36, S37, S64 and S77 (Appendix C).
In the second test two timing versions were the same as in part I, namely NAT: natural timing structure, and NEW: timing module with changed parameters (Figures 5a and 5b).

Because the female diphones were nearly twice as long as the male diphones, the parameters in the duration module that could be manipulated interactively were all divided by two as compared to the default values used for the synthesis of the male voice stimuli. The parameters of the duration

module used to generate the NEW version of the female voice were:

- first part: 35 %,
- second part: 40 %,
- accented long vowels: + 10 %,
- accented short vowels: + 10 %,
- last long vowel: + 10 %,
- last short vowel: + 10 %,
- last schwa: + 10 %,
- function words: – 15 %.

Important: The parameters of the non-interactive duration rules (like the one for lengthening of voiced consonants after lengthened short vowels) remained unaltered!

The third version had phoneme durations of 40 % of the phoneme durations in the raw diphones (see Figure 5c). (For the male voice this had been 80 %.) All versions had nearly the same total durations (approx. 2 seconds (appendix D).

Right after having completed the first part of the experiment the subjects participated in the second test. Before it started they listened to ten sentences spoken by the synthesized female diphone voice.
The three versions of the five sentences were randomized like the sentences in the first part, resulting in 5 (sentences) x 3 (versions) x 4 (repetitions) = 60 stimuli. The duration of the test was approximately nine minutes. No pause was included in the second part of the experiment.
The procedure was exactly the same as in part I of the experiment: the listeners judged the acceptability of the tempo and rhythm of three versions of the five sentences. They scored on a scale that ranged from 1 to 10, 1 being the lowest and 10 the highest score.

4.5 Results female voice

They same subjects that were discarded in the first part of the experiment, were also discarded in the second test, so that there were 840 scores (14 subject x 60 items) to be analyzed. The overall mean of the raw scores was 4.26. The means and standard errors of the mean for the three timing versions are listed in Table II. (The results per sentence are listed in Appendix E.)
After conversion of the raw scores into scale values we tested the effects analogous to the method followed in part I: we performed an ANOVA for each relevant replication factor and we found the following results:

Table II: *Means and SEM's for the three versions of the sentences synthesized with the female voice diphones (14 subjects, 5 sentences, N=70).*

Version	mean score	SEM
NAT	4.18	.11
NEW	4.43	.11
RAW	4.18	.11

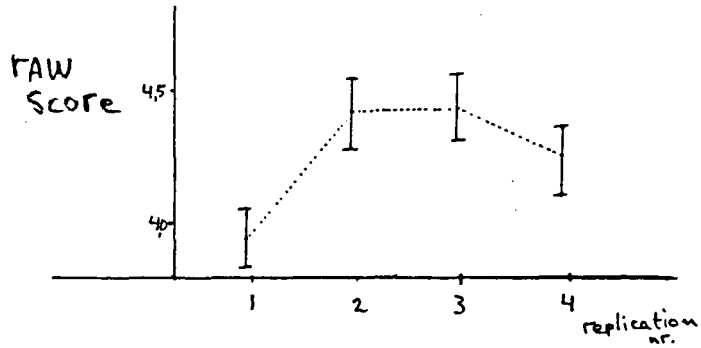

Figure 6: *Mean scores and SEM's per repetition for the female voice (14 subjects, 5 sentences, 3 versions, N=210).*

Replication factor: Sentences (5)

The version effect turned out not to be significant, and, again, the repetition effect turned out to be significant ($F_{(3,12)} = 4.20, p = .03$). There was no significant interaction effect, so that the repetition effect was the same for all versions (see Figure 6).

Replication factor: Repetitions (4)

Two effects turned out to be significant:
- the Sentence effect ($F_{(4,12)} = 38.15, p < .01$) and
- the interaction effect of versions and sentences ($F_{(8,24)} = 3.86, p < .01$).

The effect of versions turned out not to be significant.
We performed a post-hoc analysis in order to find out which version of a particular sentence was significantly more preferred. The post-hoc analysis (Scheffé's method) showed a significant Version effect for sentence S64 ('Am Zaun steht eine Regentonne') and sentence S77 ('Der Bahnhof liegt sieben Minuten entfernt'). In S64 the version RAW was judged significantly less acceptable, in S77 the version NAT was judged significantly less acceptable.

4.6 Discussion of part II

Inspection of the Repetition data (Figure 6) showed that the first time the subjects had to judge the sentences they scored much lower than the other three times, which indicates that they had to get used to the female voice.

The Sentence effect indicated that in the second test too the subjects judged some sentences to be better than others. The interaction effect of Sentences and Versions indicated that not for all sentences the same versions were preferred.

Auditory inspection of sentence versions that were judged significantly less acceptable showed that in the RAW version of S64 the diphthong /AU/ in 'Zaun' was realized too short and that /AI/ in 'eine' was too long. The problem in S77 was identical to the one in the sentence with the male voice, the last two words were too close after each other. Insertion of a short pause would improve the acceptability of the sentence.

4.7 Comparison of the results of part I and part II

In order to allow a direct comparison between the scores of the two tests we compared the raw scores and the scale values of the five sentences that were included in both tests: S16, S36, S37, S64 and S77. The mean scores for the sentences are listed in appendix F. In table III the means and SEM's for the two voices and three versions are displayed.

The total means for the two voices differed about 1.0 point in raw scores,

Table III: *Mean scores and SEM's for the three versions and the two voices (14 subjects, 5 sentences, N=70).*

	Huber		Ursula	
Version	mean score	SEM	mean score	SEM
NAT	5.15	.11	4.18	.11
NEW	5.37	.12	4.43	.11
RAW	5.33	.11	4.18	.11

the scores for the male voice being highest. After conversion of the raw scores into scale values we analyzed the data applying an ANOVA to them twice: the main effect of Voice turned out to be significant:

- replication factor Sentences: $F_{(1,4)} = 10.6, p = .03$, and
- replication factor Repetitions: $F_{(1,3)} = 20.20, p = .02$.

The interaction of Voice and Version turned out not to be significant at all ($F_{(2,8)} = .07, p = .93$).

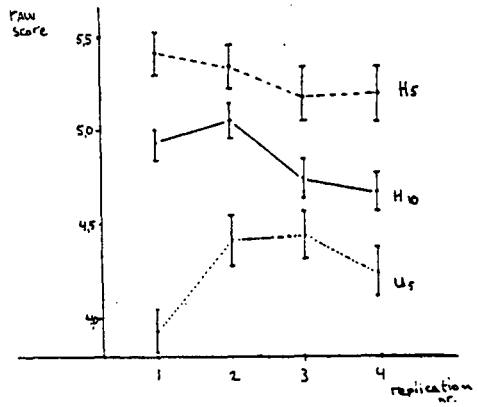

Figure 7: *Mean scores per repetition for the two voices (5 sentences, both voices, and 10 sentences, Huber; 3 versions).*

After the tests we asked the subjects to write down their judgment concerning the difference between the two voices. Most of the subjects judged the male voice to be more natural or agreeable. But, if we consider the Repetition effect (the farther in the test, the lower the scores) it seems that the listeners were getting tired or bored by listening to the same sentences again and again. The better judgment of the male voice seemed to be due to a time order effect. Listeners had to get used to the other (female) voice. As can be seen in Figure 7 the first repetition in the second test had lower scores than the other repetitions.

5 General discussion

There are several reasons why listeners might not judge timing versions of sentences to be different. Furthermore, in general the scores were rather low.

It must be considered that speech quality of the LPC-synthesized diphones was not extremely high, which might have made it difficult for listeners to abstract from overall quality in order to judge timing phenomena which were rather subtle.

Another possibility is that differences in timing were too small to be judged consistently when stimuli were presented separately instead of pairwise. In a previous test it was shown that listeners were able to detect small differences in sentence timing, the stimuli being presented pairwise. But it becomes clear that detecting timing differences is much easier than judging them.

Still another reason could be that the version with phoneme durations set to 80 % of the value produced by concatenating raw (i.e. non-adjusted) diphones, was rather acceptable. In another experiment it was shown that the natural timing version was clearly preferred above a version with phoneme durations of 100 % of the value after straightforward diphone concatenation. It could have been the case that listeners judged overall

speed of synthesized sentences, preferring the faster ones.
It might also be that timing alignment caused too much speeding up of the diphone speech considering the quality of LPC-synthesized speech. This might have caused that the somewhat slower 80 % version was judged better (it is not clear at all what cues subjects used to base their judgments upon; perhaps they judged the overall quality of the sentences instead on the timing of the sentences). This might hold for the female voices especially, since the raw diphones were shortened to some 35 or 40 % of the duration after straightforward concatenation. It resulted in very fast amplitude fluctuations, making the voice sound more curt.
We expect that longer sentences would yield larger differences between the three versions than the short sentences that were included in the test. The reason why we had to choose these particular sentences was that timing aligned versions of these sentences were available. They would not have been available for longer sentences, since they were not present in the Sotschek-corpus.

Our suggestion for further timing experiments is that relatively long sentences might be used, synthesized with a high overall speech quality. Some consideration must also be given to the method of Close Copy timing stylization. Lengthening or shortening of phoneme durations was performed after the sentence was provided with an intonation contour. Manipulation of the label values thus influenced the timing of F_0-changes. From a very recent experiment it was concluded that differences in the timing of pitch movements resulted in differently perceived (vowel) durations (see for example Rump, 1991). It was also concluded that listeners were rather sensitive to differences in the timing of pitch movements because the prominence of vowels was affected by a change in timing of pitch movements. Differences in the timing of pitch movements might also explain the results of a previous experiment (Adriaens and Rump, 1990): listeners seemed to be able to distinguish between very small differences in sentence timing versions, but they might have been distinguished differences in prominence of some phonemes.
Therefore, perceptual relevancy of peaks and valleys in the timing contour might have been artifacts: neglection of a peak may have caused a perceptually relevant difference in F_0-timing (thus affecting the intonation of the sentences), but not a perceptually relevant difference in sentence timing. The duration regularities based on the method of Close Copy timing stylization were supposed to result in perceptually relevant duration rules.
Furthermore, very short pauses that were present in the original material might have been deleted or taken within phonemes during the segmentation of the speech. We suppose that deletion of these very short pauses (a few ms in duration) might have caused that even the NAT version was

judged to be not very acceptable to the listeners. Insertion of small pauses in the appropriate positions (that are to find out!) might improve the speech quality of synthetic speech very much.

6 Conclusion

Our main conclusion is that during the perceptual evaluation of the timing modules none of the three timing versions of the sentences was judged as best. This held for both kinds of diphones, male voice and female voice. More and longer sentences might give results that are more clear cut, but also the duration rules might be improved.

Bibliography

- Adriaens-Porzig, U. and Rump, H.H.; 1990
  Das Stilisieren von Timing-Konturen.
  IPO Internal Report 762.
- Coile, B. van; 1990
  Duurregels in spraaksynthese.
  IN: Reader COLSAS 1990, Item 13, IPO Internal Report 765.
- Edwards, A.L.; 1957
  Techniques of Attitude Scale Construction.
  Appleton-Century-Crofts, Inc., New York.
- Kaiki, N.; Takeda, K. and Sagisaka, Y.; 1990
  The control of segmental duration in speech synthesis using linguistic properties.
  IN: Proceedings "ESCA Workshop on Speech Synthesis", Autrans, France, 165-168.
- Pijper, J.R. de; 1983
  Modelling British English intonation, an analysis by resynthesis of British English intonation.
  Doctoral dissertation, University of Utrecht.
- Rump, H.H.; 1991
  Timing of Pitch Movements and Perceived Vowel Duration.
  IPO Internal Report 812.
- Santen, J.P.H. van; 1990
  Deriving text-to-speech durations from natural speech.
  IN: Proceedings "ESCA Workshop on Speech Synthesis", Autrans, France, 157-160.
- Vogten, L.L.M.; 1988
  Standaardiseren van deelfoneemduren in Zelle-difonen.
  IPO Internal Report 676.

Appendices

Appendix A

Klatt-table

In this table the absolute durations of German phonemes are represented (in ms). In the first column the phoneme is given. In the second and third column the minimal and the maximal durations are listed. In the fourth column an example of the phoneme is given in a German word.

PHONEME	INH.DUR	MIN.DUR	EXAMPLE
I	95	95	Ich
IE	110	110	IHre
AE	95	95	Elf
EH	110	110	Eben
E	25	25	stundE
A	95	95	April
AH	130	130	Aber
O	95	95	Onkel
OH	110	110	Oder
U	95	95	Und
UH	110	110	UHr
UE	95	95	bUEro
UEH	110	110	UEbel
OE	95	95	OEffentlich
OEH	110	110	OEfen
AI	110	110	EIfrig
AU	110	110	AUch
EU	110	110	EUre
P	105	105	Pater
T	95	95	TaT
K	105	105	Kater
B	95	95	Bad
D	95	95	Da
G	95	95	Gabel
F	120	120	Fahrrad
S	120	120	daS
X	120	120	aCH
CH	120	120	iCH
SCH	120	120	SCHon
Z	120	120	Sehen
M	80	80	Mutter
N	80	80	Natur
NQ	80	80	juNQe
W	60	60	Wahr
Y	110	110	Ja
H	60	60	Hoffen
L	80	80	Lanq
R	80	80	Rot
SI	250	250	¡silence¿
GS	10	10	¡glottal stop¿
EL	25	25	edEL
EM	25	25	jedEM
EN	25	25	jedEN
ER	25	25	jedER

Appendix B

Phonetic feature table

In this table German phonemes are listed together with a matrix representing their fonetic features. In the last column examples of the phonemes in German words are given.

German phonemes

Phoneme features:

Phon:	Cons	Segm	Voic	Syll	Liqu	Glid	Stri	Plos	Nasa	Aspi	Long	Sjwa	Voka	Examples:
I	-	+	+	+	-	-	-	-	-	-	-	-	+	Ich
IE	-	+	+	+	-	-	-	-	-	-	+	-	+	IHre
AE	-	+	+	+	-	-	-	-	-	-	-	-	+	Elf
AEH	-	+	+	+	-	-	-	-	-	-	+	-	+	kAEse
AEQ	-	+	+	+	-	-	-	-	+	-	+	-	+	tEInt
EH	-	+	+	+	-	-	-	-	-	-	+	-	+	Eben
E	-	+	+	+	-	-	-	-	-	-	-	+	+	stundE
A	-	+	+	+	-	-	-	-	-	-	-	-	+	April
AH	-	+	+	+	-	-	-	-	-	-	+	-	+	Aber
AQ	-	+	+	+	-	-	-	-	+	-	+	-	+	gourmaND
O	-	+	+	+	-	-	-	-	-	-	-	-	+	onkel
OH	-	+	+	+	-	-	-	-	-	-	+	-	+	Oder
OQ	-	+	+	+	-	-	-	-	+	-	+	-	+	bon
U	-	+	+	+	-	-	-	-	-	-	-	-	+	Und
UH	-	+	+	+	-	-	-	-	-	-	+	-	+	UHr
UE	-	+	+	+	-	-	-	-	-	-	-	-	+	bUEro
UEH	-	+	+	+	-	-	-	-	-	-	+	-	+	UEbel
OE	-	+	+	+	-	-	-	-	-	-	-	-	+	OEffentlich
OEH	-	+	+	+	-	-	-	-	-	-	+	-	+	OEfen
OEQ	-	+	+	+	-	-	-	-	+	-	+	-	+	parfUM
AI	-	+	+	+	-	-	-	-	-	-	+	-	+	EIfrig
AU	-	+	+	+	-	-	-	-	-	-	+	-	+	AUch
EU	-	+	+	+	-	-	-	-	-	-	+	-	+	EUre
P	+	+	-	-	-	-	-	+	-	-	-	-	-	Pater
T	+	+	-	-	-	-	-	+	-	-	-	-	-	TaT
K	+	+	-	-	-	-	-	+	-	-	-	-	-	Kater
B	+	+	+	-	-	-	-	+	-	-	-	-	-	Bad
D	+	+	+	-	-	-	-	+	-	-	-	-	-	Da
G	+	+	+	-	-	-	-	+	-	-	-	-	-	Gabel
F	+	+	-	-	-	-	+	-	-	-	-	-	-	Fahrrad
S	+	+	-	-	-	-	+	-	-	-	-	-	-	daS
X	+	+	-	-	-	-	+	-	-	-	-	-	-	aCH
CH	+	+	-	-	-	-	+	-	-	-	-	-	-	iCH
SCH	+	+	-	-	-	-	+	-	-	-	-	-	-	SCHon
ZCH	+	+	+	-	-	-	+	-	-	-	-	-	-	Genie
Z	+	+	+	-	-	-	+	-	-	-	-	-	-	Sehen
M	+	+	+	-	-	-	-	-	+	-	-	-	-	Mutter
N	+	+	+	-	-	-	-	-	+	-	-	-	-	Natur
NQ	+	+	+	-	-	-	-	-	+	-	-	-	-	juNQe
W	+	+	+	-	-	+	-	-	-	-	-	-	-	Wahr
Y	+	+	+	-	-	+	-	-	-	-	-	-	-	Ja
H	+	+	+	-	-	-	-	-	-	-	-	-	-	Hoffen
L	+	+	+	-	+	-	-	-	-	-	-	-	-	Laqg
R	+	+	+	-	+	-	-	-	-	-	-	-	-	Rot
SI	-	-	-	-	-	-	-	-	-	-	-	-	-	<silence>
GS	-	-	-	-	-	-	-	-	-	-	-	-	-	<glottal stop>
EL	+	+	+	+	+	-	-	-	-	-	-	+	-	edEL
EM	+	+	+	+	-	-	-	-	+	-	-	+	-	jedEM
EN	+	+	+	+	-	-	-	-	+	-	-	+	-	jedEN
ER	-	+	+	+	-	-	-	-	-	-	-	+	-	jedER

empty phoneme : SI

Appendix C

Tested sentences from the Sotschek-corpus (orthographic)

- *S16[3]: Wer möchte keinen Kuchen? ("Who would not want cake?")
- S26: Im Geschäft stehen viele Leute. ("In the shop there are many people").
- S35: Vater will sich eine Pfeiffe anzünden. ("Daddy wants to light his pipe").
- *S36: Seine Frau macht ein trauriges Gesicht. ("His wife is looking sad").
- *S37: Du solltest weniger rauchen. ("You should smoke less").
- S38: Die Ärzte sind damit gar nicht einverstanden. ("The doctors do not at all agree with it").
- *S64: Am Zaun steht eine Regentonne. ("At the fence there is a waterbutt").
- S71: Rückt die Stühle an den Tisch. ("Pull the chairs to the table").
- *S77: Der Bahnhof liegt sieben Minuten entfernt. ("The railway station is seven minutes away").
- S100: Das war jetzt aber ein schöner Tag. ("That was a lovely day now").

Tested sentences from the Sotschek-corpus in phonemic transcription

Backslashes indicate function words, quotes indicate sentence accents.

- *S16: W\EHER MOECHTE K'AINEN KUHXEN?
- S26: GS\IM GESCH'AEFT SCHTEHEN F'IELE L'EUTE.
- S35: F'AHTER W\IL Z\ICH GS\AINE PF'AIFE GSANTSUENDEN.
- *S36: Z\EINE FR'AU MAXT GS\AIN TR'AURIGES GEZ'ICHT.
- *S37: D\UH Z\OLTEST W'EHNIGER RAUXEN.
- S38: D\IE GS'AERTSTE Z\INT D\AHMIT G'AHER N\ICHT GSAINFAERSCHTANDEN.
- *S64: GS\AM TS'AUN SCHTEHT GS\AINE R'EHGENTONE.
- S71: RUEKT D\IE SCHT'UEHLE GS\AN D\EHN T'ISCH.
- *S77: D\EHER B'AHNHOHF LIEKT Z'IEBEN MINUHTEN GSAENTF'AERNT.
- S100: D'AS W\AHER YAETST GS\AHBER GS\AIN SCH'OEHNER TAHK.

[3]The sentences marked by an asterisk (*) are also tested synthesized with Ursula-diphones.

Appendix D

Durations of the synthesized sentences

Durations of the three versions of the test sentences (ms), for both voices. (NAT: with phoneme durations of the natural spoken sentences, NEW: timing module, RAW: 80 % of the duration of the sentences synthesized with standardized diphones).

VOICE:	HUBER			URSULA		
version:	NAT	NEW	RAW	NAT	NEW	RAW
S16	1.20	1.43	1.36	1.20	1.46	1.36
S26	1.66	1.51	1.56	-	-	-
S35	1.89	1.87	1.95	-	-	-
S36	1.82	1.87	2.02	1.82	2.13	2.20
S37	1.43	1.23	1.36	1.46	1.36	1.46
S38	2.20	2.33	2.48	-	-	-
S64	1.59	1.69	1.77	1.59	1.66	1.72
S71	1.31	1.31	1.41	-	-	-
S77	2.05	2.18	2.25	2.05	2.30	2.30
S100	1.72	1.74	1.87	-	-	-

Appendix E

Means of the raw scores of the ten Huber sentences

Sentence	Version			
number	NAT	NEW	RAW	Mean
S16	4.2	4.3	4.8	4.4
S26	5.6	5.2	4.6	5.1
S35	5.5	5.3	4.8	5.2
S36	4.7	4.7	4.5	4.6
S37	5.2	4.3	4.5	4.7
S38	5.1	3.6	4.1	4.3
S64	6.7	7.0	6.8	6.9
S71	4.5	3.5	3.6	3.8
S77	4.9	6.5	6.1	5.8
S100	4.0	3.3	3.6	3.6
Mean	5.0	4.8	4.7	4.8

Mean raw scores for the 5 sentences synthesized with the female diphones (3 versions)

Sentence	Version			
number	NAT	NEW	RAW	Mean
S16	3.1	3.7	3.2	3.3
S36	3.6	3.6	3.6	3.6
S37	4.6	4.3	4.5	4.4
S64	5.5	5.4	4.3	5.0
S77	4.0	5.4	4.3	5.0
Mean	4.2	4.4	4.2	4.3

Appendix F

Comparison of the means for both voices

	Huber	Ursula
Sentence	mean score	mean score
S16	4.4	3.3
S36	4.6	3.6
S37	4.7	4.4
S64	6.9	5.0
S77	5.8	4.8
Total	5.3	4.3

Appendix G

Score forms

TIMINGTEST 2

Hallo Testteilnehmer,

Willkommen zu diesem Perzeptionsexperiment. Dies ist ein Test mit deutscher synthetischer Sprache. Es wird Ihnen eine Anzahl von Sätzen vorgespielt. Bitte beurteilen Sie, wie Tempo und Rhythmus der Sätze Ihnen gefallen.

Die Testsätze sind mit Hilfe von sogenannten Diphonen synthetisiert. Das Experiment besteht aus zwei Tests. Im ersten Test sind die Sätze mit Hilfe von Diphonen einer Männerstimme synthetisiert, im zweiten Test mit Hilfe von Diphonen einer Frauenstimme.

Vor Beginn des ersten Tests werden Ihnen 20 Sätze vorgespielt, damit Sie sich an die Diphonsprache gewöhnen können. Bitte beurteilen Sie auch diese Sätze. Jeder Satz bekommt von Ihnen eine Note auf der Skala von 1 bis 10. Ein Satz bekommt die Note 1, wenn das Satztiming Ihnen sehr schlecht gefällt. Wenn Tempo und Rhythmus im Satz Ihrer Meinung nach stimmen, bekommt der Satz die Note 10.

Nach einem Pfeifton beginnt der erste Test, der etwa 15 Minuten dauern wird. Es werden darin 120 Sätze getestet. Zur Orientierung gibt es nach 5 Sätzen jedesmal einen Pfeifton.

Anschließend an diesem Test werden Ihnen einige Sätze (die Sie noch nicht beurteilen brauchen) mit den 'weiblichen' Diphonen vorgespielt. Nach einem Pfeifton beginnt der zweite Test, in dem Ihnen 60 Stimuli vorgeapielt werden, wird etwa 8 Minuten dauern. Bitte beurteilen sie auch das Satztiming dieser Sätze auf der Skala von 1 bis 10. Wenn es noch Undeutlichkeiten gibt, so fragen Sie bitte.

Viel Erfolg und vielen Dank für Ihre Mitarbeit.

1 = schlecht 10 = gut.

Beispiele:

Wer möchte keinen Kuchen?	1 2 3 4 5 6 7 8 9 10
Im Geschäft stehen viele Leute.	1 2 3 4 5 6 7 8 9 10
Vater will sich eine Pfeiffe anzünden.	1 2 3 4 5 6 7 8 9 10
Seine Frau macht ein trauriges gesicht.	1 2 3 4 5 6 7 8 9 10
Du solltest weniger rauchen.	1 2 3 4 5 6 7 8 9 10
Die Ärzte sind damit gar nicht einverstanden.	1 2 3 4 5 6 7 8 9 10
Am Zaun steht eine Regentonne.	1 2 3 4 5 6 7 8 9 10
Rückt die Stühle an den Tisch.	1 2 3 4 5 6 7 8 9 10
Der Bahnhof liegt sieben Minuten entfernt.	1 2 3 4 5 6 7 8 9 10
Das war jetzt aber ein schöner Tag.	1 2 3 4 5 6 7 8 9 10
Wer möchte keinen Kuchen?	1 2 3 4 5 6 7 8 9 10
Im Geschäft stehen viele Leute.	1 2 3 4 5 6 7 8 9 10
Vater will sich eine Pfeiffe anzünden.	1 2 3 4 5 6 7 8 9 10
Seine Frau macht ein trauriges gesicht.	1 2 3 4 5 6 7 8 9 10
Du solltest weniger rauchen.	1 2 3 4 5 6 7 8 9 10
Die Ärzte sind damit gar nicht einverstanden.	1 2 3 4 5 6 7 8 9 10
Am Zaun steht eine Regentonne.	1 2 3 4 5 6 7 8 9 10
Rückt die Stühle an den Tisch.	1 2 3 4 5 6 7 8 9 10
Der Bahnhof liegt sieben Minuten entfernt.	1 2 3 4 5 6 7 8 9 10
Das war jetzt aber ein schöner Tag.	1 2 3 4 5 6 7 8 9 10

Erster Test: Huber

1 = schlecht 10 = gut.

1. 1 2 3 4 5 6 7 8 9 10	26. 1 2 3 4 5 6 7 8 9 10	51. 1 2 3 4 5 6 7 8 9 10
2. 1 . . . 5 10	27. 1 . . . 5 10	52. 1 . . . 5 10
3. 1 . . . 5 10	28. 1 . . . 5 10	53. 1 . . . 5 10
4. 1 . . . 5 10	29. 1 . . . 5 10	54. 1 . . . 5 10
5. 1 . . . 5 10	30. 1 . . . 5 10	55. 1 . . . 5 10
6. 1 2 3 4 5 6 7 8 9 10	31. 1 2 3 4 5 6 7 8 9 10	56. 1 2 3 4 5 6 7 8 9 10
7. 1 . . . 5 10	32. 1 . . . 5 10	57. 1 . . . 5 10
8. 1 . . . 5 10	33. 1 . . . 5 10	58. 1 . . . 5 10
9. 1 . . . 5 10	34. 1 . . . 5 10	59. 1 . . . 5 10
10. 1 . . . 5 10	35. 1 . . . 5 10	60. 1 . . . 5 10
11. 1 2 3 4 5 6 7 8 9 10	36. 1 2 3 4 5 6 7 8 9 10	61. 1 2 3 4 5 6 7 8 9 10
12. 1 . . . 5 10	37. 1 . . . 5 10	62. 1 . . . 5 10
13. 1 . . . 5 10	38. 1 . . . 5 10	63. 1 . . . 5 10
14. 1 . . . 5 10	39. 1 . . . 5 10	64. 1 . . . 5 10
15. 1 . . . 5 10	40. 1 . . . 5 10	65. 1 . . . 5 10
16. 1 2 3 4 5 6 7 8 9 10	41. 1 2 3 4 5 6 7 8 9 10	66. 1 2 3 4 5 6 7 8 9 10
17. 1 . . . 5 10	42. 1 . . . 5 10	67. 1 . . . 5 10
18. 1 . . . 5 10	43. 1 . . . 5 10	68. 1 . . . 5 10
19. 1 . . . 5 10	44. 1 . . . 5 10	69. 1 . . . 5 10
20. 1 . . . 5 10	45. 1 . . . 5 10	70. 1 . . . 5 10
21. 1 2 3 4 5 6 7 8 9 10	46. 1 2 3 4 5 6 7 8 9 10	71. 1 2 3 4 5 6 7 8 9 10
22. 1 . . . 5 10	47. 1 . . . 5 10	72. 1 . . . 5 10
23. 1 . . . 5 10	48. 1 . . . 5 10	73. 1 . . . 5 10
24. 1 . . . 5 10	49. 1 . . . 5 10	74. 1 . . . 5 10
25. 1 . . . 5 10	50. 1 . . . 5 10	75. 1 . . . 5 10

1 = schlecht 10 = gut.

76. 1 2 3 4 5 6 7 8 9 10	91. 1 2 3 4 5 6 7 8 9 10	106. 1 2 3 4 5 6 7 8 9 10
77. 1 . . . 5 10	92. 1 . . . 5 10	107. 1 . . . 5 10
78. 1 . . . 5 10	93. 1 . . . 5 10	108. 1 . . . 5 10
79. 1 . . . 5 10	94. 1 . . . 5 10	109. 1 . . . 5 10
80. 1 . . . 5 10	95. 1 . . . 5 10	110. 1 . . . 5 10
81. 1 2 3 4 5 6 7 8 9 10	96. 1 2 3 4 5 6 7 8 9 10	111. 1 2 3 4 5 6 7 8 9 10
82. 1 . . . 5 10	97. 1 . . . 5 10	112. 1 . . . 5 10
83. 1 . . . 5 10	98. 1 . . . 5 10	113. 1 . . . 5 10
84. 1 . . . 5 10	99. 1 . . . 5 10	114. 1 . . . 5 10
85. 1 . . . 5 10	100. 1 . . . 5 10	115. 1 . . . 5 10
86. 1 2 3 4 5 6 7 8 9 10	101. 1 2 3 4 5 6 7 8 9 10	116. 1 2 3 4 5 6 7 8 9 10
87. 1 . . . 5 10	102. 1 . . . 5 10	117. 1 . . . 5 10
88. 1 . . . 5 10	103. 1 . . . 5 10	118. 1 . . . 5 10
89. 1 . . . 5 10	104. 1 . . . 5 10	119. 1 . . . 5 10
90. 1 . . . 5 10	105. 1 . . . 5 10	120. 1 . . . 5 10

Zweiter Test: Ursula

1 = schlecht 10 = gut.

1. 1 2 3 4 5 6 7 8 9 10	21. 1 2 3 4 5 6 7 8 9 10	41. 1 2 3 4 5 6 7 8 9 10
2. 1 . . . 5 10	22. 1 . . . 5 10	42. 1 . . . 5 10
3. 1 . . . 5 10	23. 1 . . . 5 10	43. 1 . . . 5 10
4. 1 . . . 5 10	24. 1 . . . 5 10	44. 1 . . . 5 10
5. 1 . . . 5 10	25. 1 . . . 5 10	45. 1 . . . 5 10
6. 1 2 3 4 5 6 7 8 9 10	26. 1 2 3 4 5 6 7 8 9 10	46. 1 2 3 4 5 6 7 8 9 10
7. 1 . . . 5 10	27. 1 . . . 5 10	47. 1 . . . 5 10
8. 1 . . . 5 10	28. 1 . . . 5 10	48. 1 . . . 5 10
9. 1 . . . 5 10	29. 1 . . . 5 10	49. 1 . . . 5 10
10. 1 . . . 5 10	30. 1 . . . 5 10	50. 1 . . . 5 10
11. 1 2 3 4 5 6 7 8 9 10	31. 1 2 3 4 5 6 7 8 9 10	51. 1 2 3 4 5 6 7 8 9 10
12. 1 . . . 5 10	32. 1 . . . 5 10	52. 1 . . . 5 10
13. 1 . . . 5 10	33. 1 . . . 5 10	53. 1 . . . 5 10
14. 1 . . . 5 10	34. 1 . . . 5 10	54. 1 . . . 5 10
15. 1 . . . 5 10	35. 1 . . . 5 10	55. 1 . . . 5 10
16. 1 2 3 4 5 6 7 8 9 10	36. 1 2 3 4 5 6 7 8 9 10	56. 1 2 3 4 5 6 7 8 9 10
17. 1 . . . 5 10	37. 1 . . . 5 10	57. 1 . . . 5 10
18. 1 . . . 5 10	38. 1 . . . 5 10	58. 1 . . . 5 10
19. 1 . . . 5 10	39. 1 . . . 5 10	59. 1 . . . 5 10
20. 1 . . . 5 10	40. 1 . . . 5 10	60. 1 . . . 5 10