DESIGN, EVALUATION AND ACQUISITION OF A SPEECH DATABASE FOR GERMAN SYNTHESIS-BY-CONCATENATION

V. Kraft[1], J.R. Andrews[2]

[1)]Institut für allgemeine Elektrotechnik und Akustik, Ruhr-Universität Bochum, Germany

[2)] Department of Electrical & Electronical Engineering, University of Nottingham, UK

ABSTRACT - This paper presents a systematic approach for the definition of a speech database to be used for parametric or non-parametric speech synthesis for German. Considering coarticulation and practical aspects, the demisyllable (DS) is chosen as the basic unit. Improvements are achieved by adding vowel-to-vowel-diphones and frequent words to the inventory. Describing the whole process of definition, recording and processing of the speech elements including the interface to the TTS-system, an account is given on the experience gained on the way to high-quality speech synthesis .

INTRODUCTION

During the last 18 months an inventory of speech segments was set up in the course of a joint British-German research project. The creation of the database was intended for different aims:

- As a unit inventory for time-domain speech synthesis-by-concatenation,
- for parametric coding of the speech segments to drive a formant synthesizer, and
- to be used for the investigation of coarticulation phenomena.

In the following sections the creation of this database will be described in detail, starting with the specification and design of the inventory. After explanation of our data acquisition and processing techniques the corresponding interface from the text-to-speech system to the synthesizer will be presented. Starting from our experiences made during the project, the last chapter before coming to the conclusions will mention possible improvements which can lead to a better quality of the data or can speed up the process of building a database.

The whole procedure has to be carried out carefully since an adverse specification or a quality degradation in one step cannot be compensated for in a following step. Under the time-consuming tasks of such a scheme, it is hoped by the authors that the aspects mentioned in this paper will help the reader to find a suitable specification for their own data acquisition.

A survey of the successive steps of the project is given in figure 1.

REQUIREMENTS OF A HIGH-QUALITY SPEECH SYNTHESIS SYSTEM

A text-to-speech system, which converts automatically any written text information into spoken information in the form of speech signals, can be divided into two processing steps:

- First comes the symbolic processing, which analyses the input text to gather as much information as possible about the underlying linguistic structure. Different kinds of linguistic knowledge, represented by rules or dictionaries, are needed here to create a phonetic and prosodic description of the utterance.
- Subsequently operates a voice generation system, that converts these discrete phonetic and prosodic symbols into a continuous speech signal.

This paper focuses on the voice generation system, also termed a speech synthesizer. The synthesizer itself can follow a parametric or non-parametric strategy: While the former method uses a source model to describe the excitation signal parametrically, the latter contains no explicit description of the source. A parametric synthesizer is mainly based on rules, which determine the parameter change over the time, whereas a non-parametric system works more on data, which are manipulated, concatenated and - if necessary - decoded into an output signal. Both strategies have their own pros and cons:

From the phonetic-scientific point of view, a rule-based synthesizer is the more attractive approach because it provides full control over the mechanism of speech generation. The price to pay for this is to have all acoustic/phonetic knowledge available. But this calls for an appropriate model which goes far beyond the well-known source-filter-model on the one hand, but allows to use a clearly arranged parameter set on the other hand. One still unsolved topic of research is the modelling of the

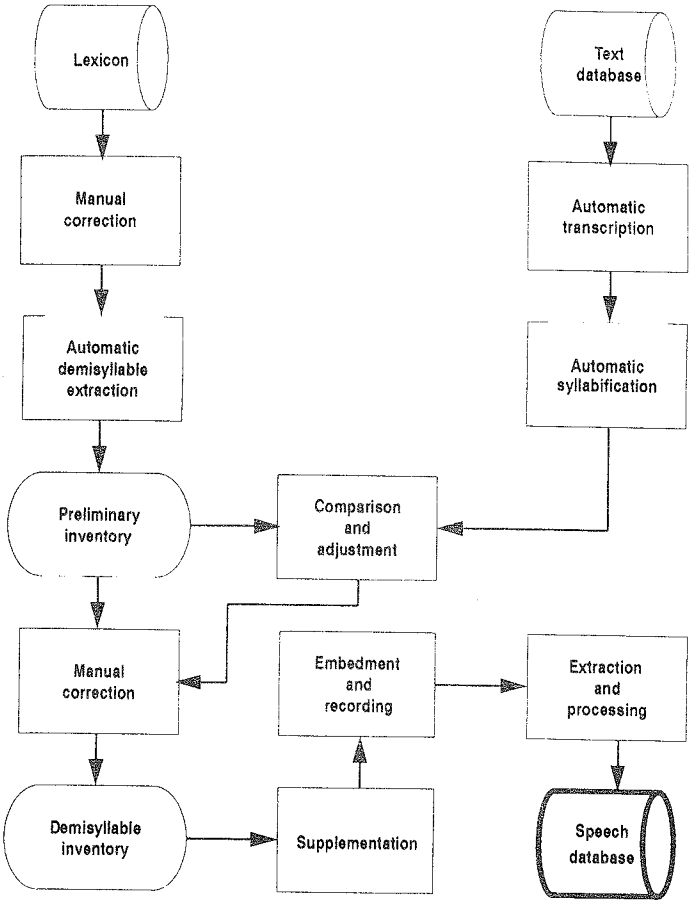

Figure 1. Survey of the whole database generation process

excitation source giving high naturalness. Further powerful tools and a suitable database are needed for analysis of natural speech to support the segmental rule development. The advantage of a parametric approach, namely the possible flexibility and variability of the speech generation process, conflicts with the complexity required for modelling natural speech - if even possible.

While a parametric synthesizer demands for computation power, this requirement is replaced by need of memory for a more data-based non-parametric synthesis. Here all phonetic rules are - invisible but perfectly - contained in the 'pre-compiled' speech elements which set up a so-called unit-inventory. The independent manipulation of the prosodic parameters intonation, duration and loudness can be performed by pitch-synchronous overlap-add- (PSOLA-) techniques (MOULINES and CHARPENTIER 1990), which preserve a maximum degree of naturalness. The coarticulation process inside the segments is perfect - because it comes from natural speech - whereas the second crux to handle is the unit concatenation in the time domain. The points where the segments are chained up in a new context to form an arbitrary utterance are possible sources for discontinuities. This fact requires that unit boundaries coincide with coarticulation barriers as much as possible and that the frequency of concatenation points will be minimal, which leads to a similar structure described in the following chapter. Instead of rule-development like in the former approach, the main tasks for building a data-based synthesizer are the definition and acquisition of the inventory, which need their own efficient tools combined with phonetic knowledge. For applications which demand different or even adaptive voice characteristics the problem of speaker transformation has to be solved in the future.

The goal of all speech synthesis research should be a system with guaranteed segmental intelligibility (important for dialog systems) combined with a maximum degree of acceptance (which in turn is composed of naturalness, ease of listening, etc. as needed, for example, for reading machines). Additionally an ideal system should offer flexible features considering the range of languages, speaker characteristics and possible applications together with a flexible and friendly user interface.

THE UNIT INVENTORY: DESIGN AND EVALUATION

Choice of the unit structure

Looking for speech elements which contain coarticulation, caused by context-dependent articulatory smoothing effects, as much as possible and which are also suitable to build a finite inventory from the technical point of view, you will at last come to the syllable. Words as the next larger units cannot be taken into account since there exists an open and further growing number of words for each language. On the other side, phones (or allophones) as the smallest segments are useful in rule-based synthesis only, since having only one (or more) steady-state target parameter-sets means that all time-variant qualities have to be modelled by rule. Switching to diphones, containing the transition from one phoneme to the next, would solve the problem to a large extend, but does not take into account coarticulation reaching beyond the next neighbouring phone nor intersyllabic allophonic variations. Since these effects are important especially for German, syllables have to be considered as the ideal basic elements. After the successful use in speech recognition, they were first recommended for speech synthesis by FUJIMURA (1976).

A survey of the existing phonetic or phonological theories about the speech-syllable term is beyond the scope of this paper (see ORTMANN 1980 for an overview for German). The fact that every child begins to stammer on the basis of syllables before speaking rhymed verses some years later shows that the not clearly defined and often intuitively used term 'syllable' has its right to exist. Most difficulties arise by the determination of syllable boundaries of more-syllabic words.

The framework of all phonologically possible syllables for German can be described by C_i-V-C_f, with C_i representing the initial consonant cluster containing 0 to 3 consonants, V for the nucleus containing one vowel or diphthong and C_f for the final consonant cluster containing 0 to 5 consonants. DETTWEILER (1983) counts 160.000 theoretically possible syllables for German, which calls for a gigantic memory, not only because of the number but also because of the large size of syllables compared to diphones. However, this last point does not allow to skip the rarely used elements, since this would cause a clear perceptible error. The fact that the nucleus acts as a coarticulation barrier in most cases advises to split all syllables within the nucleus into an initial demisyllable (IDS) and a final demisyllable (FDS). As a result both the number and the size of the elements will be reduced considerably, for the price of one additional concatenation point located in the center of a phone where a suitable interpolation method has to be applied.

Choice of the phonetic representation

For the phonemic description of the speech elements the standardised SAM Phonetic Alphabet SAMPA (ESPRIT-SAM 1992) has been chosen because of its computer applicability and multi-lingual compatibility. The number of demisyllables is nearly proportionally influenced by the set of syllabic nuclei, because every additional vowel has to be combined with all possible consonant clusters to yield the corresponding IDS and FDS. We included the following vowels into a minimal set of nuclei:

Short, open vowels:	/I, E, O, U, Y/
Short, closed vowels:	/a, 9/
Long, open vowels:	/E:/
Long, closed vowels:	/i:, e:, a:, o:, u:, y:, 2:/
Schwa and syllabic /6/	/@, =6/

The short closed /i, e/ can be generated from the long vowels /i:, e:/ by changing the quantity by rule. Additionally the three diphthongs /aI, aU, OY/ have been included into the recordings for the investigation, where the IDS can be replaced by the

IDS containing the onset to the first vowel of the diphthong only.

The consonant clusters can be composed of the plosives /p, b, t, d, g, b/, the fricatives /f, v, s, z, S, Z, C, j, x, h/, the nasals /m, n, N/, the liquids /l, R/ and the /6/ (coloured <r>). Also an empty consonant cluster /_/, preceding or following the vowel, has to be taken into account.

Definition of a complete inventory for German

A simple method to determine a (nearly) complete inventory is to combine all initial and finite consonant clusters with all vowels which would leed to an overshoot solution, since the number of really existing combinations decreases rapidly with the size of the consonant cluster. This fact was analysed systematically to take the full advantage of phonotactic constraints in minimising the inventory (see also Fig. 1).

First a lexicon was chosen for the basis of the analysis, which contained 17.000 entries on grapheme and phoneme level. The latter level was already segmented into speech syllables. After removing the foreign entries from the lexicon (but keeping some specials like /kOm-pju-t=6/), all founded DS were automatically extracted and grouped into CV, CCV, CCCV, VC, VCC, VCCC, VCCCC and VCCCCC classes. Then a list of nearly 70.000 grapheme words was automatically transcribed and then segmented into speech syllables using our TTS system SyRUB-4 (see below). Now all words that could not be composed from the DS found in the lexicon were checked: If an error occured during the transcription or the determination of syllable boundaries these errors were corrected manually before a second check against the lexicon. In the cases where the TTS system worked correctly but there were still missing elements, the inventory was adjusted. This process was continued until all (correctly) segmented phoneme words could be assembled from the preliminary DS-inventory.

The result was a corpus of 1080 IDS and 804 FDS. All possible C_i-V-C_f-combinations can be grouped into a matrix representation separated for each class, like the following example for the V-CCC case (1/0: combination exists / does not exist):

VKKKK	a:	a	E:	E	U	Y
6tst	1	0	1	0	0	0
RCts	0	0	0	0	1	0
Rnst	0	0	0	1	0	0
Rpst	0	0	0	1	0	0
Rtst	0	1	0	1	0	1
Rkts	0	1	0	0	0	0
lpst	0	0	0	1	0	0
mpft	0	1	0	1	1	0
nfts	0	0	0	0	1	0
ntst	0	1	0	1	0	0

This homogenous DS-inventory was then supplemented for two different reasons:

In the case that both the initial and final consonant cluster between two succeeding syllable nuclei are empty (for example like the German words <Trauer> /tRaU_-_=6/ or <Bio> /bi_-_O/, where '-' stands for a syllable boundary, '_' for an empty consonant cluster), instead of a relatively simple segment concatenation by rule in these cases an interpolation is required. Considering a non-parametric synthesis system, this will cause some problems because the boundaries have to be interpolated in the signal domain. Therefore nearly 160 vowel-to-vowel transitions (VV-diphones) have been added to the inventory, which make any interpolation by rule unnecessary.

A second supplementation was carried out to decrease the frequency of concatenation points: The first 120 most frequent German words were added to the inventory as entire elements. As a consequence the number of concatenation points is reduced considerably since roughly every other word can be synthesized as a whole. Since most of them occur as function words in unstressed positions, this can simplify the prosodic modification if the words are already unstressed (not too long and not too loud) in the original recording. The improvement is that these words are fluently synthesized without any perceptible degradation, so that the full concentration of the listener can be directed to the words of the utterance bearing more information.

DATABASE ACQUISITION

Still a controversial question is the ideal choice of the reference speaker, who has to fulfil the following conditions:

- intelligible and pleasant voice quality,
- time-invariant pronunciation,
- full control over the prosodic features and
- the capability to read aloud from phonemic descriptions, if required in different languages.

Fortunately we could engage a professional broadcasting speaker who met all conditions mentioned above for the English, French and German language. It is important that the speaker reads every sentence without emotions - more like a 'robot' than a human. Therefore an actor is likely to be not the ideal cast.

For the recording material the speech elements in focus were embedded into roughly 2000 meaningless carrier sentences, including some dummies at both the beginning and end of the list. Due to the embedment it was easier for the speaker to keep his fundamental frequency and durational patterns nearly constant during the whole session. Separately for IDS, FDS and VV the sentences were grouped according to the vowel. All speech elements were positioned in three-syllabic carrier words in such a manner that they got a secondary stress, i.e. halfway between unstressed and primary stress as a good basis for prosodic manipulation in both directions. We used the following context for embedment, where the first two sentences were already proposed by PORTELE (1990):

IDS:	/ das vE:R@ kRi-t@lal g@maxt /
FDS:	/ das Sto:sg@b-ORf ist o:n@ zIn /
VV:	/ das Sto:b-Ua ta:t o:n@ zIn /

The complete recording session took 4 hours, which was near the limit where the session would has to be better split up into several smaller ones. In future sessions we would avoid having so much /s/, /z/ and /x/-sounds included in the carrier material, which overtax the speaker to a large extend.

The recording of the material was done in an anechoic environment using high-quality equipment. Of great advantage is a direct visual feedback of volume, fundamental frequency and speed to the speaker. As a source for the later determination of fundamental period marks ('pitch'-marks) a laryngograph was used for the recording of the glottal excitation on a second channel. For the later synchronisation with the corresponding utterance, the delay line of the speech signal relative to the laryngogram was calculated by adding the mouth-microphone distance to the estimated vocal tract length.

PROCESSING OF THE SPEECH ELEMENTS

A time-consuming process was the course manual segmentation and cutting of the elements. To speed up the process of creating an additional inventory, we are currently working on an automatic procedure for identification and cutting segments out of carrier phrases, in the case that the phonetic transcription is known. The same method should be used for automatic segmentation of the speech signal on the phoneme level, which is needed for phoneme-based duration control as well as for the anchoring of intonation patterns.

Special attention has to be given to determine the concatenation points within a syllable nucleus, since they are very sensitive for perceptible discontinuities. Instead of context-independent segmentation of all elements beforehand, we store the optimal concatenation points in an IDS-FDS matrix in pairs for each nucleus to choose the ideal context-dependent concatenation points at run-time. An investigation of a suitable and synthesizer-dependent distance measure on which this matrix is based is on the way. Only a full optimised segmentation can avoid the need for parametric interpolation and therefore a possible loss of quality.

INTERFACING THE TTS-SYSTEM

Driving a speech synthesizer described above needs a TTS system with a corresponding output interface. A basic requirement is the automatic segmentation of phonemic syllables, which is also needed for lexical stress assignment in TTS. In addition to the phonemic representation, morphological information is taken into account. As the morphological decomposition is performed referring to orthography, a close synchronisation is necessary for syllable segmentation.

According to the definition of the phonemic syllable there must be just one boundary between two succeeding syllabic phonemes. If there are no other phonemes in between, the phoneme boundary is also the syllable boundary. Otherwise, the rules of ORTMANN (1980) are applied:

- Beginnings of roots and prefixes indicate syllable boundaries.

- A single consonant between two syllable nuclei: If the left-hand syllabic phoneme is a short vowel and the consonant is not a plosive, the boundary is put in the middle of the consonant, otherwise at its onset.
- Two syllabic phonemes are separated by a consonant cluster: Generally, the syllable boundary coincides with the onset of the last consonant. In case of a cluster-final /ks/, /ps/, /ts/ or /kv/, the boundary lies at the onset of the last but one consonant. If a cluster-final and morph-initial /R/ or /l/ is preceded by a plosive, /S/, /v/ or /f/, the syllable onset is also put left of the last but one member of the consonant cluster.

After that all syllable boundaries have been determined, a verification is made based on the available elements in the unit inventory. In case of an element missing in the database, three way outs are checked:

- As a first trial, the problem can be solved by moving the syllable boundary phoneme-by-phoneme towards the right and left neighbouring nucleus.
- If not successful, some common substitution rules can be applied (i.e., /g/ in V-C elements can be replaced by /k/).
- As the last resort alternatives can be chosen in the (seldom) case of a missing element.

The duration and intonation interface should not be considered here.

FURTHER IMPROVEMENTS AND CONCLUSIONS

Today the best synthesis quality can be obtained by non-parametric methods. Thereto this paper discussed some relevant questions concerning design and methodology. Some improvements still have do be done:

- An exception handling (proper names!) has to be implemented, if the utterance cannot be composed from the DS-inventory. One possible solution are diphones, which can be located within the DS, so they do not require additional memory.
- The discontinuity problem caused by concatenation has to be solved by better segmentation and interpolation techniques.
- Methods of further reduction of the inventory have to be examined, like the FDS-splitting in the HADIFIX-system (PORTELE, 1990), diphthong substitution, etc.
- The whole process has to be more automised, so that a speech synthesizer can be easily extended by additional voices or languages.

Further research in speech processing and speech perception is needed to answer these and more questions, before a speech synthesizer can be applied in broad applications to the user's satisfaction.

ACKNOWLEDGEMENTS

The authors wish to thank the DAAD and the British Council for funding this project under the Academic-Research-Collaboration scheme. Additionally we want to thank Prof. Dr.-Ing. J. Blauert and Dr. K.M. Curtis who initiated and supervised the project.

REFERENCES

Dettweiler, H. (1983), *Automatische Sprachsynthese deutscher Wörter mit Hilfe von silbenorientierten Segmenten*, Dissertation, University of Munich, Germany.

ESPRIT-SAM Multi-lingual Speech Input/Output Assessment, Methodology and Standardisation (1992), *Standard Computer-Compatible Transcription*, Final Report SAM-UCL-037.

Fujimura, O. (1976), *Sylable as the Unit of speech synthesis*, Int. Rep., AT&T Bell Labs, Murray Hill, NJ, USA.

Hess, W. J. (1992), *Speech Synthesis - A Solved Problem?*, EUSIPCO-92, Bruessels, Belgium, p. 37-46.

Moulines, E. & Charpentier, F. (1990), *Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones*, Speech Communication 9 (5/6), 453-467.

Ortmann, W.D. (1980), *Sprechsilben im Deutschen*, Goethe-Institut, Munich, Germany.

Portele, T. et al. (1990), *HADIFIX: a system for German speech synthesis based on demisyllables, diphones, and suffixes*, Proceedings of the ESCA-Workshop on Speech Synthesis, Autrans (F), p. 161-164.