

The SVOX Text-to-Speech System

Beat Pfister, September 1995

This document gives an overview of the current state of the SVOX text-to-speech (TTS) system, which has been developed at TIK/ETHZ.

Conceptual considerations

Correctly reading aloud a text requires far more than word pronunciation knowledge. In general, complete understanding of the semantic content and of the pragmatic context is necessary, regardless of whether a human or a machine is reading (cf. [HP87]).

Since the automatic semantic and pragmatic analysis of arbitrary texts is far from being reality, general purpose TTS systems can rely on simpler knowledge only, in particular on syntactic and morphological language knowledge. Although, from a linguistic point of view, this way of proceeding is very rough, it delivers very often quite acceptable results, mainly because there is a strong relation between the syntactic structure and the semantic and pragmatic content of a text.

SVOX has been designed as a general purpose TTS system. Its linguistic processing includes the morphological and syntactic levels only.¹ In order to achieve a system concept that is well adapted to the particularities of each processing level, and thus can be con-

sidered as systematic solution rather than a fast non-durable approach, the following design guidelines have been adopted:

- a) Speaker-independent parts of the TTS system like syntactic and morphological analysis, syllable accentuation, phrasing, etc. are strictly separated from speaker-dependent parts, i.e., from all speech signal related processing. According to this guideline the whole system splits in two main parts, and it is necessary to know, what information has to flow from the speaker-independent to the speaker-dependent part.
- b) For this interdisciplinary task, the type of knowledge representation and processing has to be adapted to the type of knowledge available for each subtask. Thus, e.g., morphological knowledge is adequately represented in the form of a morpheme lexicon and associated word production rules, where as phone duration knowledge is rather suited to be statistically modeled.
- c) The speech signal production should be based on the concatenation of natural speech segments and on a method for pitch and duration modification with minimum impairment of the segmental speech quality.²

¹Most of the TTS systems that have been realized so far, include only a dramatically simplified syntactic processing or even no syntax analysis at all, and therefore produce prosodically very unnatural sounding speech.

²It is well-known, that other speech production methods like the formant or the articulatory model potentially have the same or possibly even better speech quality, but a satisfying control of all dynamic parameters has hardly ever been achieved.

System description

In compliance with the above considerations, the SVOX text-to-speech synthesis system consists of two main parts, called transcription and phono-acoustical model, as shown in figure 1.

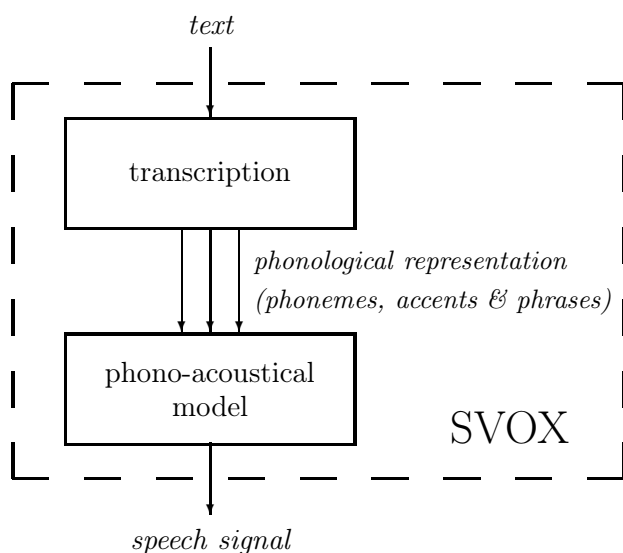


Figure 1: The two main components of the SVOX text-to-speech system: The transcription and the phono-acoustical model that reflect the speaker-independent and the speaker-dependent part, resp.

The *transcription* includes the text analysis and the phonological generation, i.e., all speaker- or voice-independent parts of the system. The output of the transcription consists of a purely abstract representation of the speech signal generated from the input text. This abstract or phonological representation is generated from each input sentence and comprises the respective phoneme string, the accent level per syllable, and the phrase boundaries (position, type and strength).

The *phono-acoustical model* includes all the speaker-dependent components that are required to generate from the phonological representation an appropriate speech signal. The two following sections provide more de-

tailed information about the transcription and the phono-acoustical model.

The transcription

The transcription comprises a morphological and syntactic analyzer for German words and sentences. This analyzer has been realized as a chart parser that uses a word and a sentence grammar together with a full-form and a morpheme lexicon. The lexicons include morpho-syntactic information and phoneme strings. By applying so-called two-level rules, lexical graphemic and phonemic forms are transformed into the corresponding surface forms and vice versa. The word and sentence grammars are specified in the DCG (definite-clause grammar) formalism. From the syntactic structure produced by the parser, a rule-based subsystem determines prosodic phrase boundaries and accents.

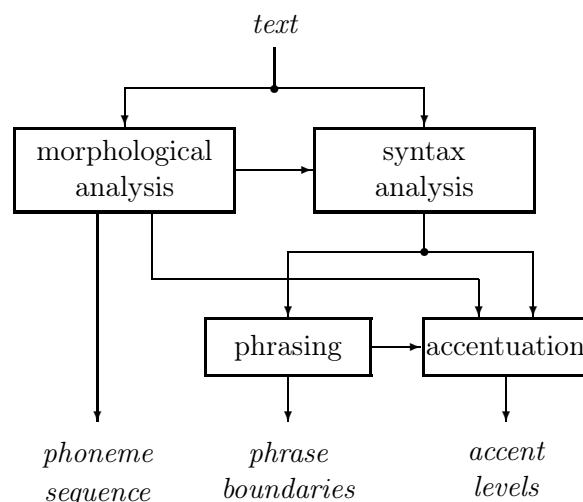


Figure 2: The transcription component of the SVOX text-to-speech system.

For words that cannot be found in the full-form lexicon and that are also rejected by the regular morphological analysis, there exists a fall-back procedure based on a DCG formulation of possible phonotactic and graphotactic structures of German words. Phonemic

transcriptions of unknown words can thus be found by parsing these words according to the phonotactic grammar and with a lexicon of German graphemic and phonemic consonant and vowel clusters (cf. [Tra95]).

The phono-acoustical model

The two main tasks of the phono-acoustical model are the prosody control and the speech signal generation, which is done by diphone concatenation. Based on the phonological representation of a sentence, the prosody control predicts phone and pause duration values and fundamental frequency contours by means of neural networks. These neural networks have been trained with a lot of examples of natural prosody (further information is available e.g. in [Hub90], [Tra93], and [Rie95]).

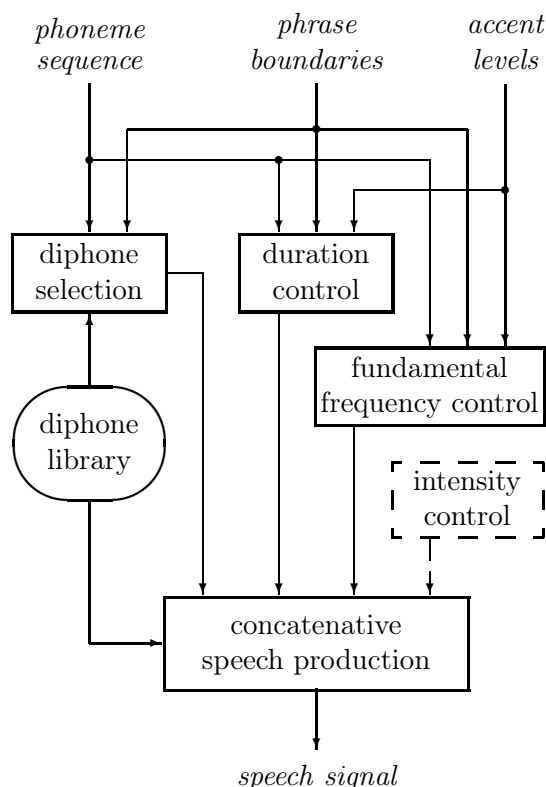


Figure 3: The phono-acoustical model of the SVOX text-to-speech system (the current version lacks an intensity control; the intrinsic intensity of the diphones is used instead).

The speech production component modifies the diphones, which have been extracted from natural speech (cf. [Kae85]), according to the predicted duration values and the fundamental frequency contours (in the current version of SVOX the intrinsic intensity of the diphones is used). The prosodic modifications of the diphones are imposed either with the LPC (linear predictive coding) method or the newer TD-PSOLA (time-domain pitch-synchronous overlap-add) algorithm.

State of the project

The prototype version of SVOX includes, apart from the intensity control, all components. It is a complete software solution that can be run in real-time on a powerful workstation. In its current form, it is rather adapted to further research purposes than to minimum memory and computing power requirements. Thus, currently SVOX is preferably used in workstation-based applications like centralized automatic information systems.

Although SVOX is probably the best TTS system for German in terms of speech quality, there are very many applications, where it is not directly usable. In many cases one of the problems is associated with the correct pronunciation of proper names. This pronunciation requires not only lexical knowledge but also to some extent multilingual capabilities, because e.g. French proper names cannot be synthesized via concatenation of German diphones.

An additional weakness of the current version of the SVOX system is its lack of appropriate prosody control for dialogue situations, which may differ substantially from the normal text reading style.

In some applications, however, the problem is not primarily associated to the speech synthesis, but rather results from the manner

how the different system components communicate with each other. Thus, systems that include semantic knowledge processing should not send pure text to the speech synthesizer, but rather some semantic representation of the utterances to be synthesized (concept-to-speech). This applies in particular to dialogue systems and machine translation systems with audio output.

All these problems are due to the preliminary state of the SVOX system. Its overall concept, however, is perfectly well suited to be extended in direction to future requirements like multilinguality and concept-to-speech operation.

Future research

The aim of the future research work is primarily to overcome the above mentioned limitations. In particular, high priority has been assigned to the incorporation of some multilingual capabilities for French (phonetic and morphological levels) and to the adaptation for concept-to-speech operation.

With the multilinguality to be reached in the near future, the SVOX system will be able to appropriately articulate German sentences with embedded French words or short expressions. For that purpose, primarily two parts of the system have to be extended: The morphological analyzer needs an additional morpheme lexicon for French and the corresponding word grammar. The diphone library has to be extended in order to include all French diphones. Additionally, some minor modifications of the fundamental frequency and the duration control will be required.

For the adaptation of the system for concept-to-speech (input of semantic concepts or trees) instead of text-to-speech, the syntax analysis component will be replaced by a much simpler component. Its task is reduced roughly

to grapheme-to-phoneme conversion. An important feature of concept input is, that focal and contrastive accents are additional components of the input information, which must be appropriately transformed into prosodic features. This requires some extensions of the fundamental frequency control and the duration control.

(for further information or a demonstration of the SVOX system, please refer to: <http://www.tik.ee.ethz.ch/cgi-bin/w3svox>)

Bibliography

- [HP87] K. Huber, B. Pfister, et al. Sprachsynthese ab Text. In *Analyse und Synthese gesprochener Sprache*. Tagungsband Nr. 9 der Gesellschaft für linguistische Datenverarbeitung, S. 26–33. Georg Olms Verlag, 1987.
- [Hub90] K. Huber. A statistical model of duration control for speech synthesis. In *Proc. of the EUSIPCO, Barcelona*, September 1990.
- [Kae85] H. Kaeslin. *Systematische Gewinnung und Verkettung von Diphonementen für die Synthese deutscher Standardsprache*. Diss. Nr. 7732, Institut für Elektronik, ETH Zürich, Januar 1985.
- [Rie95] M. Riedi. A neural network-based model of segmental duration for speech synthesis. In *Proceedings of Eurospeech'95*, pages 599–602, September 1995.
- [Tra93] C. Traber. Syntactic processing and prosody control in the SVOX TTS system for German. In *Proceedings of Eurospeech'93*, pages 2099–2102, September 1993.
- [Tra95] C. Traber. *SVOX: The Implementation of a Text-to-Speech System for German*. PhD thesis, No. 11064, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 7, ISBN 3 7281 2239 4), March 1995.