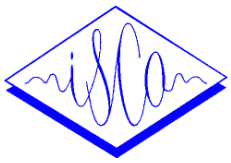

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A NEW SYSTEM FOR TEXT-TO-SPEECH CONVERSION, AND ITS APPLICATION TO SWEDISH

Mats Ljungqvist *Anders Lindström* *Kjell Gustafson‡*

Telia Promotor Infovox AB, Box 2069, S–171 02 Solna, Sweden
‡Also with the Dept. of Speech Communication, KTH, S–100 44 Stockholm, Sweden
E-mail: firstname.lastname@infovox.se

ABSTRACT

High quality text-to-speech conversion requires robust and accurate text processing, well founded modelling of phonological/phonetic processes, and sound generation with high intelligibility and naturalness. The present paper describes a new text-to-speech system under development which provides a flexible framework for integrating these processing modules. We give an overview of the system design and describe some recent advances in the related fields: text processing centered around a large lemma-based lexicon, the rule formalism and data structures of the system, and some recent work on improving voice source modelling in a formant synthesizer.

INTRODUCTION

The process of text-to-speech conversion involves many different types of processing and covers many disciplines. The present paper describes a text-to-speech system under development which provides a framework for integrating all the processing modules required for the complete text-to-speech conversion process. Although the current target language is Swedish, care has been taken to keep language independent basic mechanisms separate from language specific knowledge, in order to facilitate the development of TTS for other languages using this system.

Our original Infovox TTS system is based on RULSYS [1, 2], and over the past decade we have developed TTS systems for about a dozen languages. Many features of the RULSYS system are present in the new system. This has been designed in such a way that we can re-utilize the large amount of multi-lingual TTS technology inherent in our old system. System design concepts have also drawn on the ideas from the PROPHON system [3], e.g. rule formalisms, data structure, and the reliance on a large lexicon as the primary knowledge source for predicting transcriptions with letter-to-sound rules used as a fall-back strategy. We have also adopted the concept of a synchronized, multi-level data structure which has been explored by other researchers [8, 14]. To this framework we have added an advanced text processing stage [10], improved rule formalisms, and flexibility in the choice of sound generation model.

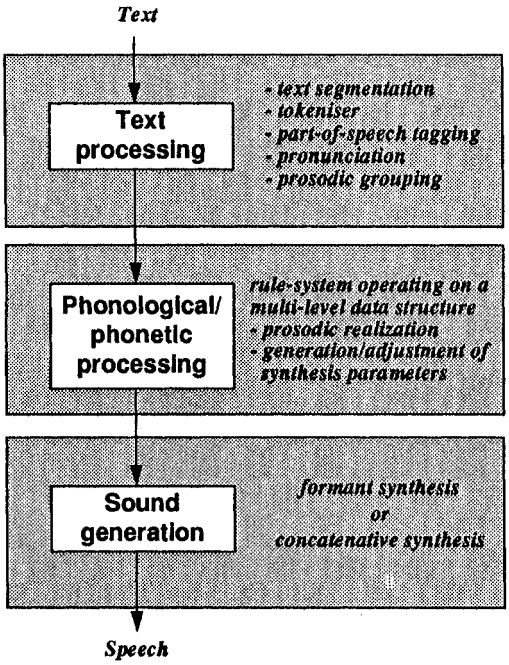

Figure 1: *System architecture for the text-to-speech conversion process.*

SYSTEM DESIGN

The system described in this paper consists of three major components: text processing, phonological/phonetic processing, and sound generation as shown in Fig. 1. They will be described in the following sections. Our main criteria in the design of the system have been:

- Modularity: for maintainability, and easy integration of new knowledge
- Flexibility: variation of voice types, speech rates etc.
- High quality synthesis: improved acoustic quality, improved text analysis

Text Processing

Text processing is carried out in a framework that supports multiple hypotheses, both in the case of segmentation into lexical units ("token hypotheses") which may

consist of multi-word expressions, and in the case of the interpretation of these units ("word hypotheses"). This process is described in more detail in [10]. The central *knowledge source* at this stage is a large lemma-based lexicon which also includes a large number of collocations (recurrent multi-word expressions), abbreviations, acronyms and proper names. If lexicon search fails, grapheme-to-phoneme rewrite rules are used to produce a phonetic transcription. Alternatively, methods based on analogy may be used. A probabilistic part-of-speech tagger achieves disambiguation of "token hypotheses" and "word hypotheses" and provides input to a module which performs grouping into prosodic words and phrases. This grouping is later used by the prosodic realization module. Number processing and application dependent processing are also carried out at this stage.

Upon leaving the text processing component all ambiguities on the text level are resolved. The result is represented by a tree-structure whose nodes correspond to syntactic and/or prosodic phrases and whose leaves correspond to word objects. The word objects contain information pertaining to the word: orthography, phonetic transcription, syllable information, lexical accent, syntactic label, etc.

Phonological/phonetic Processing
The phonological/phonetic processing module takes care of the necessary adjustments to account for assimilation, coarticulation, and reduction phenomena, as well as the realization of prosody. At this stage, a powerful rule formalism is available. While having access to the tree-structure from the text processing stage, the rules operate on a multi-level data structure consisting of several "streams": sentence, word, syllable and segment, each level having its own set of parameters, binary features and variables (Fig. 2). The same data structure and rule formalism are also available for prosodic realization where gestures in fundamental frequency (F_0), segment timing and intensity are generated.

Various models for prosody generation can easily be accommodated within this framework. For instance, the F_0 contour can be modelled in terms of "high" and "low" points with their associated interpolation functions [7] or in terms of phrase and accent commands feeding second order smoothing functions [11]. The synthesis parameters are specified as time-value-function triplets where the function may be selected from a library of functions and models. This subsegmental parameter specification is finally delivered to the sound generation module which may be either a formant synthesizer or a concatenation based synthesizer.

Rule formalisms The general rule format is of the type:

`condition → action/leftContext _ rightContext`

where `condition` specifies the focus of the rule (which could be on any level) in terms of features, phonetic symbols and/or syntactic labels, `action` specifies the required manipulation of the unit in focus, while `context`

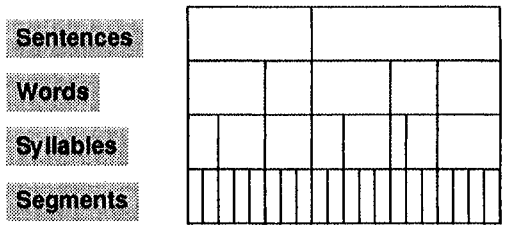

Figure 2: *Synchronized multi-level data structure of the phonological/phonetic component*

specifies the environment in which the unit in focus appears. Some of the features of the rule formalism are:

- several streams, with binary features and continuous variables in each stream,
- α, β, γ notation for compact rule writing, example: [k]-- >[α`front`]/_ [`voc`, α`front`],
- mathematical functions and logical functions,
- regular expressions: `word:[-focus]{1,3}` (at least one, not more than three non-focal words), `seg:[-voice]*` (any number of unvoiced segments),
- library of concatenation functions for synthesis parameters (cos, lin, etc.),
- local variables within a rule,
- specification of rule types and rule blocks,

Rule types Different rule types capable of reading and writing in different streams can be declared, serving the dual purpose of encouraging the developer to structure the rule base and of increasing the execution efficiency. Examples of declarations are shown below. Rules of the type *word_assignment_rules* can read the `word`, `syllable` and `segment` streams and write to the `word` stream. Rules of the type *prosody_realization_rules* can read the same streams but write to the `segment` stream.

```
rule word_assignment_rules =
[word, syll, seg] --> word

rule prosody_realization_rules =
[word, syll, seg] --> seg
```

Rule blocks Rules are split up into named blocks (modules) containing one or several rules of previously declared types. Rules can be explicitly labelled, or identified by the rule block name (when tracing and debugging, for instance). The first rule below exemplifies a stress rule assigning focal accent to the last content word, the second rule exemplifies the possible realization of focal accent for a Swedish accent-1 word as an F_0 gesture located on the syllable carrying primary stress.

```
block assignment: word_assignment_rules
{
word:[+noun | +adj | +verb] --> [+focus] /
_ word:[-noun, -adj, -verb]*
}

block realization: prosody_realization_rules
{
word:[+focus,accI] syll:[+stress, +primary]
seg:0 --> [insert F0[n] {40, 130, (cos)}]
}
```

Sound Generation

The new system can be fitted with different kinds of synthesizers. One can thus choose to use a "concatenative synthesizer" for applications requiring the naturalness offered by such synthesizers. A formant synthesizer may be used in other applications requiring its greater degree of flexibility for modelling varying speech rates, voice types etc. As a formant synthesizer we have adopted the new GLOVE synthesizer developed at KTH.

Formant Synthesizer The GLOVE synthesizer is an upgraded version of the OVE III synthesizer [9]. The new or improved features involve:

- Nasals
- Voiced fricatives
- Aspiration noise
- Voice source characteristics

The LF voice-source, [6] is a four parameter model of glottal flow derivative which enables a good match to observed waveshapes. In order to facilitate the development of text-to-speech rules, analytically and statistically based transformations of LF-parameters have recently been developed [4, 5]. The purpose is twofold, to allow a reduction of the number of parameters while retaining waveshape essentials, and to add parameters that are more closely related to phonatory mechanisms and phonetic categories than the original LF-parameters. In synthesis this set of parameters is translated back to the conventional LF-parameters which are the GLOVE input parameters.

The main pulse shape parameter of the transformed set is U_o/E_e, where U_o is the peak value of oscillatory glottal flow, and E_e is the flow derivative negative peak at the instant of excitation. The overall significance of U_o/E_e is the ratio of voice fundamental amplitude to formant amplitudes. It has been found [4, 5] that there exists a high degree of correlation between E_e and F_0 and also between U_o and E_e. A major part of the U_o and E_e variations in vowels thus follow the intonation contour.

Tentative voice source rules involve the influence of F_0, overall variation of subglottal pressure within a phrase, segmental inherent reductions of E_e and U_o due to supraglottal interaction, coarticulation effects across segment boundaries, time constants at segmental boundaries and at phrase boundaries, stress, prominence and choice of word accent. We are currently carrying out analysis and resynthesis of a corpus of speech material which is based on continuous recording of U_o and E_e from simplified inverse filtering with supplementary complete LF-parameter determinations. We are formulating rules for the automatic assignment of LF parameter values based on this work.

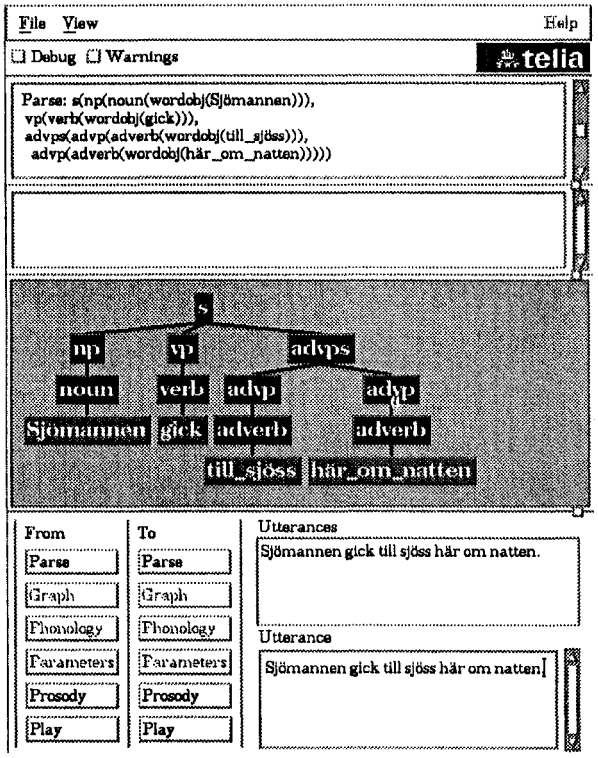

Figure 3: *The graphical interface of the text-to-speech development system*

Concatenation Based Synthesizer A version of the described system has been successfully interfaced with a concatenation based synthesizer based on the PSOLA technique [12] and the concatenation of units of arbitrary size "polyphones" [13].

Development Environment

To make the system useful as a tool for research and development it has to provide the linguist/phonetician with an environment for the development and debugging of rules and other knowledge sources. The current system implements an interactive environment through a graphical interface as shown in Fig. 3. In this environment the developer is able to interactively modify individual rules or rule blocks and study the effects of the modifications. The formant synthesizer includes a graphical parameter editor for interactive exploration of various synthesis parameter settings. Powerful debug and trace facilities are also available with features such as watchdogs, spypoints, selective printout of features, variables and synthesis parameters, and trace of the rule matching process.

DISCUSSION AND CONCLUSION

We have presented in this paper a text-to-speech system under development. The aim of the development has been to produce a system which has the flexibility of our RULSYS-based multi-lingual TTS system and is capable of utilizing the increased computing power and improved programming facilities which have become available since our original system was developed, and, not least, in which it is possible to incorporate in a structured way the various research results emerging from the international research community.

In order to achieve such a system, we have felt it to be imperative to emphasize modularity throughout the system (for easy maintainability, testing, debugging, and future extensions to the system). This paper has addressed some issues relating to each of the three main modules of the system: text processing, phonological/phonetic processing, and sound generation. Some of the benefits we are expecting to gain from the new system and its improved structural concepts, are

- increased flexibility,
- greatly increased transcription accuracy,
- enhanced lexical disambiguation ability,
- realistic prosody assignment based on syntactic and semantic processing made possible by a large lemmatized lexicon as well as by modules dedicated to syntactic parsing and semantic interpretation,
- enhanced sound quality.

Although the system described here is still under development, we are confident that it will represent a significant advance in TTS technology. The significance lies both in the practical sphere of improved, better-sounding and more accurately transcribing TTS products, but also in the potential use of the system as a research tool.

ACKNOWLEDGEMENT

We are grateful to Gunnar Fant for his encouragement and support during our work, in particular for sharing with us the results of his voice source studies, and for help in preparing this paper.

References

[1] R. Carlson and B. Granström. A text-to-speech system based entirely on rules. In *Conf. Record of the International Conference on Acoustics, Speech, and Signal Processing*, pages 686–688, Philadelphia, 1976. IEEE.

[2] R. Carlson, B. Granström, and S. Hunnicutt. A multi-language text-to-speech module. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1604–1607, Paris, 1982. IEEE.

[3] K. Ceder and B. Lyberg. The integration of linguistic levels in a text-to-speech conversion system. In *Proc. of the International Conference of Spoken Language Processing*, pages 97–100, Kobe, Japan, 1990.

[4] G. Fant, A. Kruckenberg, J. Liljencrants, and M. Båvegård. Voice source parameters in continuous speech. A progress report. In *Proc. of the Third Intl. Conf. on Spoken Language Processing*, Yokohama, 1994.

[5] G. Fant and J. Liljencrants. Data reduction of LF voice source parameters. In *Papers from the Eighth Swedish Phonetics Conference*, volume 43 of *Working Papers*, pages 62–65, Lund, Backagården, May 1994.

[6] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, (4):1–13, 1985.

[7] E. Gårding and G. Bruce. A presentation of the Lund model for Swedish intonation. volume 21 of *Working Papers*, pages 69–75. Department of Linguistics, Lund University, 1981.

[8] S. R. Hertz, J. Kadin, and K. J. Karplus. The Delta rule development system for speech synthesis from text. *Proceedings of the IEEE*, 73(11):1589–1601, 1985.

[9] J. Liljencrants. The OVE III speech synthesizer. *IEEE Trans. on Audio and Electroacoustics*, AU-16(1), 1968.

[10] A. Lindström and M. Ljungqvist. Orthographic processing *within* a speech synthesis system. In *Proc. of the Third Intl. Conf. on Spoken Language Processing*, Yokohama, 1994.

[11] M. Ljungqvist and H. Fujisaki. Generating intonation for Swedish text-to-speech conversion using a quantitative model for the F_0 contour. In *Proc. of the European Conference on Speech Technology*, volume 2, pages 873–876, Berlin, Sept. 1993. ESCA.

[12] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467, 1990.

[13] M. Rayner, H. Alshawi, I. Bretan, D. Carter, V. Digalakis, B. Gambäck, J. Kaja, J. Karlgren, B. Lyberg, S. Pulman, P. Price, and C. Samuelsson. Speech to speech translation system built from standard components. In *Proceedings of the ARPA Workshop on Human language Technology*, Plainsboro, NJ, 1993.

[14] H. van Leeuven and E. te Lindert. Speech maker: Text-to-speech synthesis based on a multi-level synchronised data structure. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 781–784, 1991.